THE THEORY OF FUNCTIONS
OF A
REAL VARIABLE
AND
THE THEORY OF FOURIER'S SERIES

VOLUME I

CAMBRIDGE UNIVERSITY PRESS
LONDON : FETTER LANE, E.C. 4

NEW YORK THE MACMILLAN CO
BOMBAY
CALCUTTA } MACMILLAN AND CO., LTD
MADRAS
TORONTO - THE MACMILLAN CO. OF
CANADA, LTD
TOKYO MARUZEN KABUSHIKI KAISHA

ALL RIGHTS RESERVED

THE THEORY OF FUNCTIONS
OF A
REAL VARIABLE
AND THE
THEORY OF FOURIER'S SERIES

BY

E. W. HOBSON, Sc.D., LL.D., F.R.S.
SADLEIRIAN PROFESSOR OF PURE MATHEMATICS, AND FELLOW OF CHRIST'S COLLEGE,
IN THE UNIVERSITY OF CAMBRIDGE

VOLUME I

THIRD EDITION
REVISED THROUGHOUT AND ENLARGED

CAMBRIDGE
AT THE UNIVERSITY PRESS
1927

First Edition (in one volume) 1907

Volume I
(Second Edition) 1921
(Third Edition) 1927

Volume II
(Second Edition) 1926

PRINTED IN GREAT BRITAIN

PREFACE TO THE THIRD EDITION

IN the preparation of this new edition of Volume I the text of the corresponding volume published in 1921 has been subjected to careful revision. Various sections have been rewritten; among these may be mentioned the sections on the Riemann-Stieltjes integral. A considerable amount of new matter has been added; and thus the volume has been lengthened by upwards of sixty pages. The additions have been made without changing the numeration of the sections, so that the references in Volume II (1926) to sections in Volume I are applicable to this new edition

The parts of the subject which were dealt with in the first five chapters of the first edition, published in 1907, are represented in the eight chapters of the first volume of this new edition With a view to greater unity of treatment of the Theory of Integration, some theorems which appeared in Chapter VI, of the first edition, have however been included in the present volume A considerable part of the Theory of Integration, in relation to series and sequences, is dealt with in Volume II, published in 1926

On controversial matters connected with the fundamentals of the Theory of Aggregates, account has been taken of the considerable diversity of opinion amongst Mathematicians, which still exists, but in general no attempt has been made to give dogmatic decisions between opposed views. In view of the delicate questions which arise as to the legitimacy and meaning of the axiom known as the Multiplicative Axiom, or as the Principle of Zermelo, the policy has been adopted of so framing the proofs of theorems as to avoid an appeal to the axiom, whenever that course appeared to be possible, in other cases, the necessity for the employment of the axiom has been pointed out.

Ample references to sources of information are given throughout, but such references do not provide the means for compiling a complete list of writings on the subject. No attempt has been made to settle questions of priority of discovery, but in this new edition the dates of publication of writings referred to have been inserted.

The corrections and additions given at the end of Volume II, published in 1926, have been embodied in this new edition of Volume I. The opportunity has been taken of giving at the end of the volume some corrections and additions to Volume II.

E. W. HOBSON

CHRIST'S COLLEGE, CAMBRIDGE.
May 11, 1927.

reduce these fundamental notions to a minimum of indefinables from which the whole theory might be deduced by means of formal logic. A slight perusal of the extremely extensive literature of the Philosophy of Arithmetic will shew that any such attempt could only have been made by entering upon a prolonged discussion of a philosophical character, wholly unsuited to a treatise of professedly mathematical complexion, and that any views expressed would have but little prospect of giving general satisfaction to logicians and philosophers. The modern theory of Real Numbers has been the subject of much criticism by philosophers and others. It has been represented that the modern extension of the notion of number to the case of irrational numbers is a sophistical attempt to obliterate the fundamental distinction between the discrete and the continuous. I venture to think that such objections consist, in large part at least, of criticisms of the current terminology of the mathematical theories, especially in respect of the extensions of the use of the word "number," and I think it probable that many of these criticisms would not survive a fair examination of the theories themselves apart from the language in which they are xpressed. An appropriate terminology, although a matter of convention, s no doubt a very important matter in relation to such fundamental natters, as it is conducive to clearness of thought; but the substance of he theories is of incomparably greater importance than the forms in which hey are expressed, and those theories may be found on examination to e essentially sound, even if their terminology be regarded as in some espects defective.

Chapter II contains an exposition of the theory of sets of points, and ncludes an account of transfinite cardinal and ordinal Arithmetic, of a omewhat simpler and less general character than will be met with in the reatment of the general theory of aggregates, in Chapter III. Students ho do not care to embark upon the discussions in Chapter III will find a udy of Chapter II amply sufficient to enable them to apply the ideas here developed in the general theory of functions. A slight account only as been given of the properties of plane sets of points. An account of the mportant recent investigations which had their origin in Jordan's theorem, at a closed curve divides plane space into two regions, would have ccupied more space than was at my disposal. This omission will be less lt than might have been the case, were not an excellent account of this bject to be found in Dr W. H. Young's treatise on the theory of sets of oints, which has appeared since this portion of the present work was inted.

In Chapter IV, there will be found a discussion of the main properties functions, in relation to continuity, discontinuity, &c., and investigaons of the properties of important classes of functions. Although the eatise is mainly one of functions of a single variable, a considerable amount of space has been devoted to the consideration of functions of two variables, not only on account of the intrinsic importance of that subject, but because no adequate consideration of the properties of functions of a single variable is possible without the use of functions of two variables, as is seen, for example, from the consideration that a function defined by means of a sequence of functions of a single variable is virtually defined as a limit of a function of two variables.

The foundations of the Integral Calculus, as based upon Riemann's definition of a definite integral, and its extensions, are discussed in Chapter v, where an account of the development of the subject from the point of view of Lebesgue's new definition of the definite Integral is also given In later parts of the book I have introduced extensions of Lebesgue's definition, to the cases of improper integrals, taken over finite or infinite domains, regarded as the limits of sequences of Lebesgue integrals.

Chapter vi is concerned with functions defined as the limits of sequences of functions, and contains an account of the principal properties of functions represented by series, and a discussion of important matters connected with the modes of convergence of series through whole intervals, or in the neighbourhood of particular points Various matters relating to the processes of the Integral Calculus, which had not been considered in Chapter v, are here dealt with, because their adequate treatment presupposes a knowledge of the theorems relating to the convergence of sequences of functions An account of the very general results recently obtained by Baire, relating to the representability of functions by means of series, will be found in this Chapter.

Chapter vii is devoted to the theory of Fourier's series. No apology is needed for the selection of this particular mode of representation of functions for full discussion in a treatise on the theory of functions of a real variable, in view of the historical relation of Fourier's series to the development of the general theory. The history of the theory of Fourier's series is exceedingly instructive, not merely from the point of view of the mathematician, but also from that of the epistemologist. I have therefore endeavoured, in my treatment of the subject, to preserve as much of the historical element as was possible in an account which should contain, in a moderate compass, not only indications of the various stages of development of the subject, but also the most recent results that have been obtained. I have made full use of the greater generality which can be introduced into many of the known results by means of the employment of the theory of integration developed by Lebesgue.

In the preparation of the work, the treatises from which I have most largely drawn information are the German edition of Dini's treatise on the subject, Stolz's *Grundzüge der Differential- und Integral-Rechnung*, Schoenflies' *Bericht* entitled *Die Entwickelung der Lehre von den Punkt-*

mannigfaltigkeiten, and the various treatises on different parts of the subject by Borel and Lebesgue. I have consulted a very large number of memoirs, articles, notes, and books, far too numerous to be here particularized. In respect to the references given throughout the book, I wish it to be understood that I have made no attempt to settle questions of priority of discovery The references given are to be regarded solely as indicating sources of information from which I have drawn, or where more detailed information on the various topics is to be found.

I owe a debt of gratitude to my friend Mr J. W. Sharpe, formerly Fellow of Gonville and Caius College, who has read with the greatest care the proofs of about two-thirds of the book Many points of difficulty I have fully discussed with him; many obscurities of expression have been removed, and many improvements in substance have been made, owing to the care he has bestowed in reading the proofs. I felt it as a great loss when, owing to a temporary failure of health, he was unable to continue his laborious work. To Dr H. F. Baker, F.R.S., Fellow of St John's College, and Cayley Lecturer in Mathematics, who has kindly read some of the earlier proofs, I owe several valuable suggestions. On several points connected with the treatment of Number in Chapter I, I have had the advantage of consulting Dr James Ward, F.B.A., Fellow of Trinity College, and Professor of Mental Philosophy and Logic in the University.

My thanks are due to the officials of the University Press for the readiness with which they have met my views, and for the care which they have bestowed upon the work connected with the printing. I desire especially to express my sense of the value of the excellent work done by the readers of the Press, to their care is due the elimination of many typographical and other blemishes which would otherwise have remained undetected.

E. W. HOBSON

CHRIST'S COLLEGE, CAMBRIDGE.
May 15, 1907.

CONTENTS

CHAPTER I

NUMBER

SECTIONS		PAGES
1	Introduction	1–3
2–4	Ordinal numbers	3–7
5	Mathematical induction	7–8
6, 7	Cardinal numbers	8–11
8–10	The operations on integral numbers	11–13
11–13	Fractional numbers	14–17
14, 15	Negative numbers, and the number zero	18–20
16, 17	Irrational numbers	20–22
18	Kronecker's scheme of arithmetization	22–23
19–22	The Dedekind theory of irrational numbers	23–27
23–28	The Cantor theory of irrational numbers	28–34
29	Convergent sequences of real numbers	35–37
30–32	The arithmetical theory of limits	37–40
33, 34	Equivalence of the definitions of Dedekind and Cantor	40–42
35	The non-existence of infinitesimals	42–43
36–38	The theory of indices	43–47
39, 40	The representation of real numbers	47–51
41, 42	The continuum of real numbers	51–54
43	The continuum given by intuition	54–55
44, 45	The straight line as a continuum	55–59

CHAPTER II

DESCRIPTIVE PROPERTIES OF SETS OF POINTS

SECTIONS		PAGES
46	Introduction	60–61
47, 48	The upper and lower boundaries of a linear set of points	61–64
49	Non-linear sets of points	64–65
50	Limiting point of a convergent sequence of intervals, or cells	65–67
51	Systems of nets	68–70
52–54	The limiting points and the derivatives of a set	70–75
55	Descriptive terminology	75–79
56	Properties of closed and open sets	79–81
57	Property of the successive derivatives of a set	81
58–60	Enumerable aggregates	81–86
61, 62	The power, or cardinal number, of an aggregate	86–88
63	The arithmetic continuum	88–91
64–66	Transfinite ordinal numbers	91–97
67	Properties of aggregates of closed sets	97–98
68	The transfinite derivatives of a set of points	98–100
69–72	Sets of intervals or cells	100–106
73–77	The Heine-Borel theorem	106–113
78, 79	The Lebesgue chain of intervals	114–116

SECTIONS		PAGES
275, 276	Oscillating continuous functions	374–377
277, 278	Properties of incrementary ratios	377–379
279–285	Properties of the derivatives of continuous functions	380–386
286	Functions with one derivative assigned	386–387
287–290	The construction of continuous functions	387–391
291–300	General properties of derivatives	391–404
301	Functions of two variables	404–405
302–306	Double and repeated limits	405–414
307, 308	The limits of monotone and quasi-monotone functions of two variables	414–417
309, 310	Partial differential coefficients	417–422
311–315	Higher partial differential coefficients	422–432
316–319	Functions defined implicitly	432–443
320, 321	Maxima and minima of a function of two variables	444–446
322–324	Properties of a function continuous with respect to each variable	447–451
325–328	The representation of a square on a linear interval	451–458

CHAPTER VI

THE RIEMANN INTEGRAL

329	Introduction	459–460
330	The Riemann integral in a linear interval	460–461
331–334	The upper and lower Riemann integrals	462–468
335	Particular cases of functions that are integrable (R)	468–470
336	Geometrical interpretation of Riemann integration	470–472
337	Properties of the definite Riemann integral	472–476
338–340	R-integrals of functions of two or more variables	476–480
341, 342	Integrable null-functions and equivalent integrals	480–482
343–349	The fundamental theorem of the Integral Calculus	482–491
350	Functions which are linear in each interval of a set	491
351	Integration by parts	491–493
352, 353	Cauchy's definition of an improper integral	493–498
354–359	Riemann integrals over an unbounded interval	498–506
360, 361	Change of the variable in a single integral	506–509
362–365	Repeated integrals	509–518
366–368	Improper double integrals	518–526
369–371	The double integral over an infinite domain	527–531
372–375	The transformation of double integrals	531–538
376–376^2	The Riemann-Stieltjes integral	538–546
377–380	The generalized Riemann-Stieltjes integral	546–559
381	The Riemann-Stieltjes integral for functions of two variables	559–561

CHAPTER VII

THE LEBESGUE INTEGRAL

382	Introduction	562
383, 384	Measurable functions	562–564
385–388	The Lebesgue integral of a measurable function	565–572
389	Other definitions of an integral	572–573
390	The L-integral as the measure of a set of points	573–575

SECTIONS PAGES
391 The *R*-integral as an *L*-integral 575
392, 393 The Lebesgue integral as a function of a set of points . . . 576–578
394 Equivalent *L*-integrals 578
395–399 Properties of the Lebesgue integral 579–584
400 The limits of a sequence of measurable functions 584–585
401–403 The derivatives of a function 585–587
404–409 Indefinite integrals 587–596
410–414 The fundamental theorem of the Integral Calculus for a Lebesgue integral 596–605
415 The total variation of an indefinite integral 605–606
416, 417 The generalized indefinite integral 607–608
418, 419^3 The indefinite integral of a function of two variables . . . 608–616
420 Integration by parts for the *L*-integral . . . 616
421–426 Mean value theorems 616–626
427–429 Repeated Lebesgue integrals 626–632
430–433 A fundamental approximation theorem 632–640
434–436 Approximate representation of an *L*-integral as a Riemann sum . 640–646
437–439 Lebesgue integrals over an infinite field 646–649
440–442 Change of the independent variable in a Lebesgue integral . . 649–656
443, 444 Harnack's definition of an integral 657–662
445–447 The Lebesgue-Stieltjes integral . . . 662–666
448, 448^1 The *RS*-integral for functions of two variables 666–668
449–452 Hellinger's integrals 669–674

CHAPTER VIII

NON-ABSOLUTELY CONVERGENT INTEGRALS

453–455 Harnack-Lebesgue integrals 675–679
456 The *HL*-integral over a finite set of intervals 679–680
457–461 The conditions for the existence of an *HL*-integral . . . 680–689
462 The second mean value theorem for an *HL*-integral . . . 689–691
463 Integration by parts for the Harnack-Lebesgue integral . . 691
464–466 The Denjoy integral 692–699
467–471 The fundamental theorem of the Integral Calculus for the Denjoy integral 699–709
472–475 Properties of the Denjoy integral 709–715
476–478 Extensions of the definition of the Denjoy integral . . . 715–719
479–482 The Young integral 719–726

CORRECTIONS AND ADDITIONS TO VOLUME II (1926) 727

LIST OF AUTHORS QUOTED IN VOLUME I 731

GENERAL INDEX TO VOLUME I 734

CHAPTER I

NUMBER

1. The operation of counting, in which the integral numbers are employed, can be carried out by a mind to which discrete objects, which may be either physical or ideal*, are presented, and which possesses certain fundamental notions which we proceed to specify.

(1) The notion of *unity*, a form under which an object is conceived when it is regarded as a single one. An object so regarded may be either of a material or of a purely abstract or ideal nature, and may be recognized, for all other purposes than that of counting, as possessing any degree of complexity. It is sufficient, in order that the object may be regarded under the form of unity, that it be so far distinct from other objects, as to be recognized at the time when it is counted, as discrete and identifiable. What external marks are necessary that an object may be so recognized as discrete, is a matter for the judgment of the mind at the time when the object is counted. The unity under which the object is apprehended is a formal or logical, rather than a natural unity, it is more or less arbitrarily attributed to the object by the mind.

(2) The notion of a *collection* or *aggregate* of objects, which is conceived of as containing more or fewer objects, or as possessing a greater or less degree of plurality. A group of objects regarded as an aggregate is conceived of, not merely as a plurality of objects to each of which unity is ascribed as in (1), but also as itself an object to which unity is ascribed when it is regarded as a single whole. The single objects of which the aggregate is composed may be spoken of as the *elements* of the aggregate; such elements need not possess any parity as regards size or any other special quality, but may be of the most diverse characters. a certain logical parity is however ascribed to them in the process of counting, in virtue of the fact that each of them is regarded as a single object. A sensibly continuous presentation cannot be regarded as an aggregate containing a plurality of elements, until the mind has recognized in it sufficiently distinct lines of division to serve the purpose of marking off distinct objects within it, the totality of which makes up the whole presentation; for instance, the history of a country could be regarded as

---

* It is held by some authors that the operation of counting is primarily applicable to physical objects only. Thus, J. S. Mill writes:—"The fact asserted in the definition of a number is a physical fact. Each of the numbers, two, three, four, etc., denotes physical phenomena, and connotes a physical property of these phenomena." See *Logic*, 9th edition, vol. II, p. 150. That objects which are not physical, can be counted, was maintained by Leibnitz and by Locke. See also Frege's *Grundlagen der Arithmetik*, Breslau, 1884, where an account is given of various views as to the nature and origin of the idea of Number.

an aggregate of distinct periods, only when sufficient salient features had been recognized in that history to warrant a judgment that periods were to be found in it, each of which had a sufficient degree of discreteness to be subsumed under the form of unity. In actual counting, the aggregate is not necessarily determinate before the counting is commenced, but becomes so when the process is completed; the notion of an aggregate is thus still necessary to the process of counting, if the process is ever to come to an end, or to be conceived of as having come to an end.

It has been held* that, when an aggregate is counted, the elements must remain distinct from one another, not disappearing or combining with each other during the process. That this condition is unnecessary may be seen, for example, by considering the case of counting breakers on the sea-shore, or that of counting the vibrations of a pendulum, thus no physical permanence, but only an ideal one, is necessary.

A discussion of the characteristics which an aggregate (not necessarily finite) must possess, in order that it may be an object of mathematical thought, will be given in Chapter IV.

(3) The notion of *order*, in virtue of which relative rank is given to each object in a collection, so that the collection becomes an ordered aggregate. In actual counting, the order is assigned to the objects during the process itself, as an order in time, and this may be done in an arbitrary manner; the order of the elements in an aggregate may, however, be assigned in a manner dependent upon their sizes, weights, or other qualities, or in accordance with their positions in space Order may, however, be regarded as an abstract conception, independent of a particular mode of ordering; for an aggregate to be an ordered one, it is necessary that in some manner or other, each element be recognized as possessing a certain rank, in virtue of which it is known as regards any two elements which may be chosen, which of them has the lower, and which the higher rank. An element is said to precede any other element of higher rank than itself.

(4) The notion of *correspondence*, which underlies the process of tallying The elements of one aggregate may be made to stand in some logical relation with those of another one, so that a definite element of one aggregate is regarded as correspondent to a definite element of another aggregate.

The correspondence may be complete, in the sense that, to every element of either aggregate there corresponds one element, and one only, of the other aggregate; or the correspondence may be incomplete, in which case one of the aggregates has one or more elements to which no elements in the other aggregate correspond. In the latter case we say that the aggregate with the superfluous element or elements contains more elements than the other aggregate, and that the latter contains fewer elements than the former.

---

* See Helmholtz's *Zahlen und Messen*, Leipzig, 1887; *Wissens. Abhandl* vol III, p 372.

A correspondence between two aggregates is defined when specifications or rules are laid down which suffice to decide which element of one aggregate corresponds to each element of the other; so that, in the case of complete correspondence, no element of either aggregate is without a corresponding one in the other

Whether, or how far, these fundamental notions of *unity*, *aggregate*, *order*, and *correspondence* should be regarded as derived empirically from experience, by a process of abstraction, or whether it must be held that they are original forms which the mind possesses prior to, and as the necessary conditions of the possibility of such experience, are questions into which it is beyond our province to enter It is certain that civilized man possesses these fundamental notions, and it is highly probable that primitive man possessed them long before the notion of abstract number had appeared in an explicit and developed form. The investigation of the origin of these notions, and their further analysis, are matters for the Psychologist and for the Philosopher Mathematical Science, as any other special science, must take its fundamental notions as data; it is concerned with the analysis of them, only so far as suffices to establish that they possess the degree of definiteness which such data must have, if they are to lie at the base of a logically ordered system.

ORDINAL NUMBERS

2. If some of the elements are removed from an ordered aggregate, the aggregate which remains is said to be a *part* of the original aggregate. It will be observed that the relative order of any two elements in the part is the same as the relative order of those elements in the original aggregate.

An ordered aggregate is said to be *finite* when it satisfies the following conditions:

(1) There is one element which has lower rank than any of the others.

(2) There is one element which has higher rank than any of the others.

(3) Every part of the aggregate which contains more than one element has an element which has higher rank than every other element in the part, and also it has an element which has lower rank than any other element in the part.

These conditions are equivalent to the statement that a finite aggregate, and also each part of it, has a first and a last element.

Every part of a finite ordered aggregate is also a finite ordered aggregate.

If M be the aggregate, and M_1 a part of it containing more than one element, then M_1 has a highest and a lowest element; also every part of

M_1, being also a part of M, has a lowest and a highest element in case it contains more than one element; therefore M_1 is itself finite.

3. Two finite ordered aggregates are said to be *similar* when they can be made completely to correspond, so that to each element of either of them there corresponds a single element of the other, and so that to any two elements P, Q of the one there correspond two elements P', Q' of the other, which have the same relation as regards rank; viz. that, if P is of lower rank than Q, then P' is of lower rank than Q', and if P is of higher rank than Q, then P' is of higher rank than Q'.

Two finite ordered aggregates which are similar are said to have the same *ordinal number.*

If each of two ordered aggregates is similar to a third, they are similar to one another. For if an element P of the first corresponds to an element R of the third, and the element Q of the second corresponds to R, it is clear that, if we make P correspond to Q, the first two aggregates are made to correspond in such a way that the relative order is preserved

It thus appears that *an ordinal number is characteristic of a class of similar ordered aggregates*

An aggregate which consists of a single element A is said to have the ordinal number one, denoted by the symbol 1. The ordinal number 1 is characteristic of every aggregate which consists of a single element.

If, to the aggregate which consists of an element A, we adjoin a new element B, and assign to B a higher rank than A, we obtain an aggregate (A, B) which has an ordinal number 2, characteristic of all aggregates which are formed in this manner; A is said to be the first element, B the second. If, to an ordered aggregate (A, B), of which the ordinal number is 2, we adjoin another element C, and regard this as having higher rank than A and B, we obtain an ordered aggregate (A, B, C), of which the ordinal number is called 3, and is characteristic of all ordered aggregates formed in this manner. Proceeding in this way, if we have formed an ordered aggregate $(A, B, C, \dots H)$, of which the ordinal number is n, and adjoin to this aggregate a new element K, we obtain a new aggregate

$$(A, B, C, \dots H, K),$$

of which the ordinal number n' is different from n.

Any ordered aggregate which is formed in the manner described is finite.

This can be proved by induction. Let us assume that M is a finite ordered aggregate: it will then be proved that (M, e), the ordered aggregate obtained by adjoining an element e of higher rank than the elements of M, is also finite. Since M has a lowest element, (M, e) has the same lowest element, also (M, e) has a highest element e. Again, if M_1 is a part of (M, e) which does not contain e, then M_1 is a part of M, and therefore has a highest and a lowest element. If M_1 is a part of (M, e) which contains e, let it be (M_2, e), where M_2 is a part of M, and therefore contains a lowest element which is also the lowest element of (M_2, e); also (M_2, e) contains a highest element e. It has thus been shewn that (M, e) satisfies the requisite conditions that it should be finite, provided M does so. The aggregates A, (A, B) are clearly finite: hence the method of induction proves that every ordered aggregate which can be formed by continually adjoining new elements to an aggregate which originally contained one element is a finite one.

Conversely, it can be shewn that *every finite ordered aggregate can be formed in the manner above described.*

Let M be a finite ordered aggregate, and let e' be its highest element, thus $M = (M_1, e')$. Now M_1, being a part of M, has a highest element e'', thus $M_1 = (M_2, e'')$, or $M = (M_2, e'', e')$. Proceeding in this manner, if we do not reach an aggregate M_r which contains a single element only, there must exist a part $(\ldots\ldots e''', e'', e', e)$ of M which has no element of lowest rank. But this is impossible, since M is by hypothesis finite, and therefore contains no part without a lowest element It has thus been shewn that M can be reduced, in the manner indicated, to an aggregate with a single element. and conversely, starting with this latter aggregate, M is obtained by adjoining to it successively new elements.

A finite ordered aggregate is not similar to any part of itself.

This theorem may also be proved by induction. For if we assume that the finite ordered aggregate M is not similar to any part of itself, it can be shewn that the same holds for (M, e). If possible let M_1 be a part of (M, e) which is similar to (M, e); then if M_1 contains e, it must be of the form (M_2, e), and if (M_2, e) is similar to (M, e), M_2 must be similar to M, which is contrary to the hypothesis that M contains no part similar to itself. If M_1 does not contain e, it must be of the form (M_2, f), where f is the element which corresponds to e in (M, e); in this case again M_2 is similar to M, and is a part of it; thus we have again a contradiction. The theorem holds for (A, B), and therefore generally.

It follows from this theorem that *the ordinal numbers* 1, 2, 3, ... which have been defined as the ordinal numbers of aggregates

$$(A), (A, B), (A, B, C), \ldots$$

are all different from one another, for each of these aggregates being a part of each of those which follow it, cannot be similar to any of the aggregates which follow it.

Each of the ordinal numbers is to be regarded as a unique ideal object in that it is a permanent object for thought. The relation of an ordinal number to an ordered aggregate of objects which is characterized by that number, may be illustrated by the analogy of the relation between *the colour* red and a particular red object.

4. *A simply infinite ascending aggregate*, or *simple sequence*, is an ordered aggregate which has no element of higher rank than all the others, and is such that every part which has an element of higher rank than all the other elements in that part is a finite ordered aggregate.

The fundamental assumption must be made that such an aggregate may be regarded as a definite object which possesses certain properties that can be formulated. The justification for the assumption is to be found in the fact that no contradiction arises in the theory based on it.

It follows from this definition that, in a simple sequence there is one element of lower rank than all the others; and further, that every part of the simple sequence has an element of lower rank than all the other elements in that part

A simply infinite ascending aggregate differs from a finite ordered aggregate in having no element which is of higher rank than all the other elements.

The totality of ordinal numbers forms a simply infinite ascending aggregate; these objects may be represented by a set of signs

$$\alpha,\ \beta,\ \gamma,\ \delta,\ \ldots$$

or

$$1,\ 2,\ 3,\ 4,\ \ldots$$

where it is assumed that some adequate scheme of such signs has been devised.

The order of the elements is assigned by the successive formation, as above, of aggregates having the various elements for their ordinal numbers, and it has been shewn that, if an aggregate has the ordinal number n, another aggregate having a different ordinal number n', taken to be of next higher rank than n, can be formed. There exists therefore no highest ordinal number.

Instead of using the expressions "of higher rank" and "of lower rank," it is usual to say that a number m is less than a number n, when m is of lower rank than n in the ordered aggregate of ordinal numbers, and that n is greater than m. The terms "greater" and "less" are borrowed from the language primarily applicable to the description of magnitudes: but in pure arithmetic and pure analysis generally, they are used only in the sense in which they indicate higher or lower rank, and this rank has no necessary reference to relations of magnitude or of measurable quantity.

The operation of counting a finite aggregate of objects of any kind may be conceived of as the process of putting the objects into correspondence with the elements of the aggregate of ordinal numbers, in such a way that, when any ordinal number has an element of the aggregate which corresponds to it, each of the preceding ordinal numbers also has an element which corresponds to it. The finite aggregate is usually ordered by the process itself, the ranks of the various elements being successively assigned to them as the counting proceeds. Those ordinal numbers which are employed in counting such an aggregate may be regarded as forming an aggregate which is similar to the given aggregate, as ordered by the process of counting. The last of the ordinal numbers employed in counting a finite aggregate is *the ordinal number*, or simply *the number* (Anzahl) of the ordered aggregate.

The theorem that an ordered aggregate is not similar to any of its parts holds only as regards finite aggregates. It will appear in the course of the discussion in Chapter IV that every aggregate which is not finite has parts which are similar to the whole, and this property is sometimes taken as the basis of the definition of an infinite, or transfinite, aggregate. For example, the aggregate of ordinal numbers 1, 2, 3, ... is similar to the part 2, 4, 6, ... which contains the even numbers only

MATHEMATICAL INDUCTION

5. The proofs of theorems in § 3 have been referred to as proofs by induction In its general form the principle of Mathematical Induction may be stated as follows:

If, in respect of a given simply infinite ordered aggregate, it be known (1), *that, in case any element of the aggregate possesses a certain property* P, *the element of next higher rank also possesses the property* P, (2), *that the element of lowest rank possesses the property* P; *it then follows that every element of the aggregate possesses the property* P.

The truth of the principle follows as a consequence of the properties assigned to the aggregate by means of the definition in § 4

For, let it be assumed, if possible, that there exists a part of the given aggregate A, such that the elements belonging to that part do not possess the property P. This part contains one or more elements; let M be such an element. Let us consider the finite aggregate which consists of all those elements of A which are of rank not higher than that of M. Of this finite aggregate there exists a part B, such that no element of B possesses the property P. The aggregate B contains the element M, and possibly other elements. Since B, being part of a finite aggregate, is itself a finite aggregate, it contains an element m of lower rank than all the others. By hypothesis, m cannot be the element of lowest rank in A, since the latter possesses the property P. Therefore there is in A an element m' of next lower rank than m; and m' possesses the property P. But by hypothesis, if m' possesses that property, so also does m. The assumption that A contains a part such that the elements of the part do not possess the property

P has thus been shewn to lead to a contradiction. The truth of the principle has therefore been established.

It has been maintained by Poincaré* that the principle of Mathematical Induction is a special characteristic of mathematical reasoning as distinct from that of general Logic, not being reducible to the principle of contradiction, as it essentially involves the employment of an unending chain of syllogisms. That this is the case is suggested by the form in which the principle is frequently applied in elementary mathematical treatises, viz that since the property *P* holds good for the first element, it therefore holds good for the second, and therefore for the third, etc.; and consequently, by the employment of an unending set of such syllogisms, it is inferred that the property *P* holds good for all the elements of the infinite aggregate. It has however been shewn above that it is not necessary to state the principle in this form, but that the truth of the principle follows by applying the ordinary Logic to deduce the consequences of the possession by the infinite ordered aggregate of certain properties, in accordance with the principle of contradiction.

CARDINAL NUMBERS

6. *If any finite ordered aggregate be re-ordered in any manner, the new ordered aggregate is finite, and has the same ordinal number as the original one.*

In order to prove this theorem, the following particular case will be first established:—If Q is a finite ordered aggregate, the aggregate (Q, e), obtained by adjoining to Q a new element e of higher rank than all the elements of Q, is similar to (e, Q), in which e has a lower rank than all the elements of Q. For let $Q \equiv (Q_1, f)$, and let us assume that the theorem holds for Q_1, *i.e.* that (Q_1, e) is similar to (e, Q_1); it follows, since a complete correspondence can be established between the elements of (Q_1, e) and (e, Q_1), that the same is true of the two aggregates (Q_1, e, f) and (e, Q_1, f). Now (Q_1, e, f) is similar to (Q_1, f, e), since Q_1 can be made to correspond to itself, e to f, and f to e, therefore (Q_1, f, e) is similar to (e, Q_1, f) or, (Q, e) to (e, Q), and thus the theorem holds for $Q \equiv (Q_1, f)$, provided it holds for Q_1. Now it clearly holds if Q_1 consists of a single element; hence by induction it holds for any finite ordered aggregate Q. To prove the theorem in the general case, let us assume that it is true for an aggregate M; it will then be shewn to be true for (M, e). For let an aggregate obtained by re-ordering (M, e) be (R, e, S), where either R or S may be absent; (R, e, S) is similar to (R, S, e), for R corresponds with itself, and it has been shewn above that (e, S) is similar to (S, e). Since (R, S) is by hypothesis similar to M, it follows that (R, S, e) is similar to (M, e), and therefore (R, e, S) is similar to (M, e). The theorem clearly holds for an aggregate (A, B) which contains two elements, hence by induction it holds for every finite ordered aggregate.

---

* See his work *La science et l'hypothèse*, Paris (1902), p. 19; also an article in the *Revue de Métaphysique*, vol. II (1894), p. 371. For a criticism of Poincaré's view, see an article by A. Padoa, *Proc. of the fifth annual Congress of Mathematicians*, vol. II, p. 471.

It follows from the theorem which has been established above, that. *for any aggregate which can be ordered as a finite ordered aggregate, the ordinal number is independent of the mode in which the aggregate is ordered.*

It will be found, when the generalization of ordinal numbers for non-finite aggregates is considered in Chapter IV, that this property, that the ordinal number of an aggregate is independent of the mode of ordering, is peculiar to finite aggregates.

7. Two aggregates are said to be *equivalent*, when their elements can be placed into correspondence so that to each element of either aggregate there corresponds one and only one element of the other aggregate.

It will be observed that the relation of equivalence differs from that of similarity, in that it contains no reference to order. It is clear that two aggregates which are each equivalent to a third are equivalent to one another.

An unordered aggregate is said to be *finite* when it can be so ordered that the ordered aggregate is finite in accordance with the definition given in § 2.

Two (finite) aggregates which are equivalent are said to have the same *cardinal number.*

It thus appears that *a cardinal number is characteristic of a class of equivalent aggregates.*

Each of the cardinal numbers is to be regarded as a unique ideal object; the relation of a cardinal number to a member of the class of equivalent aggregates of objects, of which it is characteristic, may be illustrated in the same manner as in § 3, in the case of the ordinal numbers

Since all similar aggregates are also equivalent, and since, in the case of a finite aggregate, the ordinal number is independent of the mode in which the aggregate is ordered, it follows that for every finite ordinal number there is a corresponding cardinal number.

The cardinal numbers of finite aggregates are denoted by the same symbols 1, 2, 3, ... as the corresponding ordinal numbers. The two kinds of numbers are not symbolically distinguished from each other, although logically they are not identical.

It will be seen in Chapter IV that this practical identity of ordinal and cardinal numbers is confined to the case of the numbers corresponding to finite aggregates, and therefore called finite numbers. The finite cardinal numbers form a simple sequence 1, 2, 3, ... similar to the sequence of finite ordinal numbers; the expressions "greater" and "less" are used in relation to two cardinal numbers in the same purely ordinal sense, denoting higher and lower rank, as in the case of ordinal numbers.

It is impossible, in a purely mathematical work, to enter into a full discussion of the nature and proper definition of number from a philosophical point of view. One view of number, which is widely held, is embodied in the definition by abstraction, in which the cardinal number[*] is regarded as the concept of an aggregate which remains when we make abstraction of the nature of the objects forming the aggregate, and of the order in which they are given; the ordinal number is then regarded as the concept obtained by making abstraction of the nature of the objects only, retaining the order[†] in which they are given in the aggregate. The view has also been maintained[‡] that a cardinal number is simply the class of all equivalent aggregates. A tendency has been exhibited amongst mathematicians[§] to regard numbers, at least for the purposes of analysis, as identical with the symbols which represent them. In accordance with this view, abstract arithmetic is set entirely adrift from the fundamental notions related to experience in which it had its origin, and it is thus reduced to a species of mechanical game played in accordance with a set of rules which, when divorced from their origin, have the appearance of being perfectly arbitrary; though it may, of course, be said that it is possible at the end of any arithmetical process to reconnect the symbols employed with the ideas which originally suggested them, and thus to interpret the results of the purely symbolical processes. Whatever view* be adopted as to the real nature of number and its place in a general scheme of thought, the assumption of the right to hypostatize numbers would appear to be an essential condition of the possibility of developing an abstract arithmetic, and consequently of the establishment of mathematical analysis in general

---

[*] This view is that of G. Cantor; see *Math Annalen*, vol XLVI (1895), p. 481, where the following definition is given:—"'Machtigkeit,' oder 'Cardinalzahl' von M nennen wir den Allgemeinbegriff welcher mit Hulfe unseres activen Denkvermogens aus der Menge M hervorgeht, dass von der Beschaffenheit ihrer verschiedenen Elemente m, und von der Ordnung ihres Gegebenseins abstrahirt wird." See also Peano, *Formulaires de Mathématiques*, 1901, § 32, 0 Note.

[†] Ordinal numbers are frequently regarded as logically prior to cardinal numbers, but this order of procedure is not a necessary one. In Dedekind's tract "Was sind und was sollen die Zahlen," Brunswick, 1887 and 1893, which has been translated into English by Prof W W. Beman, under the title "Essays on the Theory of Numbers," 1901, a detailed treatment of the subject is given, in which the notion of order is regarded as fundamental.

[‡] See B. Russell, *The Principles of Mathematics*, vol I (1903), chap XI

[§] For example see Heine, *Crelle's Journal*, vol LXXIV (1872), p. 173, where the matter is stated in the following plain form: "Ich nenne gewisse greifbare Zeichen Zahlen, sodass die Existenz dieser Zahlen also nicht in Frage steht." Again, Helmholtz appears to hold a view closely approaching the notion that Arithmetic is the art of manipulating certain signs according to certain rules of operation, he writes in *Ges Abh.* vol III, p 359, "Ich betrachte die Arithmetik oder die Lehre von den reinen Zahlen als eine auf rein psychologische Thatsachen aufgebaute Methode, durch die die folgerichtige Anwendung eines Zeichensystems (namlich der Zahlen) von unbegrenzter Ausdehnung und unbegrenzter Moglichkeit der Verfeinerung gelehrt wird." Reference may be made to an essay by A Pringsheim in the *Jahresberichte der d math. Vereinigung*, vol. VI (1899), p. 73, "Ueber den Zahl- und Grenzbegriff im Unterricht" In an article entitled "Die Du Bois-Reymond'sche Convergenz-Grenze," *Sitzungsberichte d bayer. Akad.* vol. XXVII (1897), Pringsheim speaks of numbers as "Zeichen, denen lediglich eine bestimmte Succession zukommt." See p. 326. This article contains various remarks on arithmetization, and especially a criticism of the views of P. Du Bois-Reymond. A searching criticism of the tendency to reduce Arithmetic to the formal manipulation of symbols is given in L Couturat's work *De l'infini mathématique*, Paris (1896), which contains a valuable account and discussion of theories of the philosophy of arithmetic.

THE OPERATIONS ON INTEGRAL NUMBERS

8 If two finite ordered aggregates A and B, of which the ordinal numbers are a and b respectively, are combined into a single ordered aggregate in which the elements of A have all lower rank than those of B, and in which any two elements of A, and any two elements of B, have the same relative orders as in the original aggregates, *then the ordinal number of the combined aggregate is said to be the sum of the ordinal numbers a and b*, and is denoted by $a + b$

It can be shewn that the new aggregate is a finite one, and that its ordinal number is unaltered if for A and B there be substituted aggregates which are similar to them; it thus appears that the sum $a + b$ is a finite number which depends only upon a and b.

The aggregate (A, B) has as lowest element the lowest element of A, and as highest element the highest element of B, moreover any part of (A, B) is of the form (A', B'), where A' is a part of A, and B' is a part of B; or else it has one of the forms A', B'. Since A', B' have each a lowest and a highest element, any such part of (A, B) has a lowest and a highest element. Thus (A, B) is finite.

Again, if A_1, B_1 are aggregates which are similar to A and B respectively, the elements of A may be placed in correspondence with those of A_1, and the elements of B with those of B_1; we have then a (1, 1) correspondence between the elements of (A, B) and those of (A_1, B_1); thus the ordinal number of (A, B) is the same as that of (A_1, B_1).

---

* References to the literature relating to the Philosophy of Number will be found in the article I A 1, "Grundlagen der Arithmetik," by H. Schubert, in the *Encyclopadie der mathematischen Wissenschaften*, vol. I, also in E. G Husserl's *Philosophie der Arithmetik*, vol. I, chaps 5 and 6, Halle (1891) Frege's treatises, the *Grundlagen der Arithmetik*, Breslau (1884), and the *Grundgesetze der Arithmetik*, Jena (1893 and 1903), also Whitehead and Russell's work, *Principia Mathematica*, Cambridge (1910 and later), may be here specially referred to, in connection with the relation of the foundations of Arithmetic to general Logic. The view that Number is fundamentally dependent on the notion of Time was developed by Sir W. R. Hamilton; see the *Dublin Transactions*, vol. XVII (II) (1835), "Theory of Conjugate Functions or Algebraic Couples with a Preliminary and Elementary Essay on Algebra as the Science of Pure Time", see also Helmholtz's essay *Zahlen und Messen* (1887), where the view is adopted that the axioms of Arithmetic have a relation to the intuitional form of time, similar to that which the axioms of Geometry have to the intuitional form of space.

Since (A, B) has the same ordinal number as (B, A) it follows that $a + b = b + a$, which is known as the commutative law of addition.

If a, b are the cardinal numbers of two finite aggregates A, B, then the cardinal number of the aggregate formed by combining the two aggregates into one is said to be the *sum* of a and b, and is denoted by $a + b$. That $a + b$ is a definite finite number, dependent only on a and b, follows at once from the corresponding theorem which has been proved for ordinal numbers.

The operation of finding the sum of two numbers a and b is known as the operation of addition, and it has been shewn that this operation is commutative. It should be observed that the sum of two numbers a and b cannot be determined merely by contemplating those numbers themselves as abstract concepts, but can only be defined as above, by referring to aggregates of which a and b are the numbers, and then combining those aggregates. The number of the combined aggregate is then conceived of as the result of a symbolical operation upon the numbers a and b. For example, the equation $5 + 3 = 8$ does not imply that the concept 8 is obtainable by placing the concepts 5, 3 as it were in juxtaposition, but can only be regarded as a symbolical expression of the fact that an aggregate of 5 objects together with one of 3 objects makes up an aggregate of 8 objects. Bearing this observation in mind, the numbers 1, 2, 3, ... are represented symbolically as the results of successive operations of addition, $1 + 1 = 2$, $2 + 1 = 3$, $3 + 1 = 4$, etc.; but these equations do not express definitions of the numbers 2, 3, 4, ..., since, from the concept unity taken by itself, no other concept is directly derivable.

The operation of addition can be extended by continued repetition. Thus the sum of $a, b, c, \dots k$ is a finite number represented by

$$a + b + c + \dots + k,$$

and, in particular, any number n is represented by $n = 1 + 1 + 1 + \dots + 1$. An immediate induction shews that the result of the operation of addition repeated any definite number of times is a finite number dependent only on the constituents of the summation.

The associative law of addition, $a + (b + c) = (a + b) + c$, follows from the irrelevancy of the order in which the operations are performed. This is seen from the contemplation of aggregates of which a, b, c are either the ordinal or the cardinal numbers.

9. If in a finite aggregate of which the number is b, each element be replaced by a finite aggregate of which the number is a, *the number of the new aggregate so formed is said to be the product of b by a*, and is denoted by ab. This operation is said to be that of multiplying b by a. By taking the aggregates to be ordered, it is seen at once that the new aggregate satisfies the conditions that it is finite, and that its number is unaltered by the substitution of similar aggregates of other objects for those originally employed Thus ab is a definite number dependent only on a and b.

It is clear that ab may be regarded as the sum $a + a + a + \dots$, where a occurs b times in the operation.

If the ordered aggregate of which the number is ab be re-ordered in the following manner:—take the first element of each of the aggregates of which a is the number, then the second elements of these aggregates, and so on, with lastly the ath elements of these aggregates, then we have as the result of the process an aggregate of which the number is a, and each element of which consists of an aggregate of which the number is b. the re-ordered aggregate has the number ba. It has thus been shewn that $ab = ba$, which is expressed by saying that the operation of multiplication of finite integers is commutative.

The distributive law for multiplication, $a(b + c) = ab + ac$, follows from the definition of the operation, by considering the aggregates of which a, b, c are the numbers.

An immediate induction shews that the repetition of the operation of multiplication any definite number of times gives a finite number dependent only on the numbers multiplied, and independent of the order in which the operations are performed

The result of the operation of multiplying the number a by itself is denoted by a^n, where n is the number of times a occurs in the product $a.a\ a\dots a$. From this definition the law $a^m\ a^n = a^{m+n}$ is directly deducible.

10. If the sum of two numbers a, b be denoted by c, the number a is uniquely determined when b, c are fixed, and it is then regarded as the result of the operation of subtracting b from c. The operation of subtraction is thus defined as inverse to that of addition. If $c = a + b$, a is obtained as the result of the operation denoted by $c - b$, which is such that $(c - b) + b = c$. It is obvious that the operation of subtraction of b from c is only possible in case $c > b$.

If the product of two numbers a, b be the number c, then the number a is uniquely determined when b and c are given; and a is regarded as the result of the operation of division of c by b. The operation of division so defined is inverse to that of multiplication; it is clear that the operation is only possible in case c is one of the class of numbers b, $2b$, $3b$, ...

FRACTIONAL NUMBERS

11. The operation of multiplying two integers a, b together is one which is always a possible operation, in accordance with the definition of the operation of multiplication which has been given above: the inverse operation of division is however, as we have seen, not always a possible one. This restriction upon the possibility of the operation of division suggests the introduction into Arithmetic of a new class of numbers, the rational fractions, which, when defined, shall be such that the operation of division, within the whole aggregate of integers and fractions, may be a possible one without restriction. Stated in algebraical form, the demand arises for a scheme of numbers such that the equation $ax = b$ shall always have a solution in x, where a, b are any two numbers which belong to the contemplated aggregate of numbers.

The actual use of fractional numbers arose historically from the necessities of the process of measurement of extensive magnitude, and the conception of a fraction which arises in this connection is the one which is used in ordinary life, and is made the basis of the treatment of the theory of fractions, even in recent scientific text-books. In accordance with this view, a unit of magnitude of some kind is divided into b equal parts, and a of these parts are taken; the resulting magnitude is then denoted by the fraction a/b.

This notion of the essential nature of a fraction, dependent as it is upon the notions of a *unit*, and of the divisibility of such unit into *equal parts*, is incompatible with the modern view that Mathematical Analysis should be developed upon the basis of a Pure Arithmetic, quite independently of all notions connected with the measurement of extensive magnitude. The modern tendency known as Arithmetization manifests itself in the construction of theories of Number and of the operations involving numbers, which depend entirely upon the conceptions connected with the process of counting; measurement being regarded as a process foreign to Pure Arithmetic. The process of counting is an exact one; whereas measurement can in practice only be carried out with a greater or less degree of approximation, and can only ideally be made an exact process. Pure Arithmetic is made the basis of Analysis, not only in accordance with the general principle that the fundamental conceptions of a branch of science should be irreducible to simpler conceptions, but also because the theory of ideally exact measurement has peculiar difficulties of its own. Our essentially inexact intuitions of spatial, temporal, or other magnitudes, necessitate a process of idealization in which the objects of perception are replaced by ideal objects subject to an exact scheme of definitions and postulates, in order that an exact science of measurement may be possible. The view is at present held by the majority of mathematicians that the nature of the abstract continuum, and that of a limit, are capable of exact formulation only in the language of a Pure Arithmetic; and that this science must therefore be developed upon an independent basis before it can be applied to the elucidation of the conceptions requisite for an abstract theory of continuous magnitude. The theory of measurement is, in accordance with this view, regarded as an application, and not as part of the basis, of Mathematical Analysis.

12. By those writers who are under the influence of the modern arithmetizing tendency, the traditional non-arithmetical definition of a fraction has been abandoned, and in its place a formal definition has been substituted, in which the fraction is regarded as an association of a pair of integers. The associated integers are regarded as making a single object, and laws of combination of these objects are then postulated.

If a, b are two integers, a new number (a, b), or in ordinary notation $\frac{a}{b}$, is formed by the association of a and b, the new number being defined to be such as to satisfy the following conditions:

(1) (a, b) is regarded as ordinally greater, equal to, or less than (c, d), according as ad is greater, equal to, or less than bc. The expressions greater, equal to, or less than, are here used, not in their primitive sense as referring to magnitude, but in the sense in which we have used them in the case of integers, as assigning relative order to the numbers.

(2) $(a, 1)$ is defined as equal to a; thus if $b = 1$, the association is regarded as equivalent to the integer a. Taking (1) in conjunction with this postulate, the new numbers have their orders assigned, not only relatively to one another, but relatively also to the integral numbers; so that the whole aggregate of integers and fractions is ordered, in the sense that, of two given numbers, it can always be said which has the higher rank.

(3) The addition of two fractional numbers is defined by

$$(a, b) + (c, d) = (ad + bc, bd).$$

(4) The multiplication of fractional numbers is defined by

$$(a, b) \times (c, d) = (ac, bd).$$

(5) The use of a fraction as an index is defined by the postulate

$$x^{(a, b)} \times x^{(c, d)} = x^{(a, b) + (c, d)},$$

where x is any number, either integral or fractional. The symbol $x^{(a, b)}$ is to be interpreted subject to this postulate, in case such interpretation is possible.

It will be observed that, in the case $b = 1$, $d = 1$, the above definitions are consistent with those which have been adopted in the case of integral numbers; and thus the new numbers, together with the integers, form an aggregate with uniform laws of operations. It is easily seen that the operations with the new numbers satisfy the commutative, associative, and distributive laws. The inverse operation of division is now one which is always possible within the domain of the numbers; thus

$$(a, b) \div (c, d) = (ad, bc).$$

The inverse operation of subtraction, $(a, b) - (c, d) = (ad - bc, bd)$, is only possible if $(a, b) > (c, d)$.

The association of a pair of integers is a "number" in quite a different sense from that in which the cardinal and ordinal numbers, hitherto discussed, are numbers. The justification of the extension of the term "number" to the fractions lies in the fact that a consistent scheme of operations can be imposed upon them, of which the laws are in agreement with those which hold for operations which involve integers only.

13. The scheme which has been above indicated suffices for a formal definition and logical development of the properties of fractions, but it is subject to the objection that it is of an arbitrary character, indeed it is not easy to see why the particular laws of operations have been postulated, except as suggested by the traditional non-arithmetical conception of a fraction.

To remedy this defect, a view of the nature of a fraction will be here given which relates the fraction with the process of counting, in such a manner that fractional and integral numbers have similar relations to that process. It will appear that the laws of combination given above naturally follow from this mode of regarding the fraction, with the exception of (5), which is however immediately suggested by the rule for integral indices.

Consider an aggregate of b objects, and out of these b objects pick out any a ($\leqq b$) of them. If we regard these a objects not only as single objects of number a, but also as belonging to an aggregate whose number is b, we may denote the a objects by (a, b), where their number a is associated with the cardinal number b of the aggregate to which they belong. This process being independent of the particular aggregate used, the abstract fraction (a, b) is related to this process in an analogous manner to that in which the number b is related to the process of counting an aggregate whose cardinal number is b. Thus the fraction (a, b), or a/b, is characteristic of an aggregate of a objects each of which belongs to an aggregate of b objects. The extension of the definition to the case $a > b$ is clear when we observe that it is unessential that the a objects taken should all belong to one and the same aggregate of b objects; it is sufficient that each of them be regarded as essentially belonging to *some* aggregate of cardinal number b. In accordance with this view, a fraction, say $3/5$, is characteristic of any three things each of which belongs to an aggregate of five things, *i.e.* $3/5$ means 3 out of 5. That the three things taken out of five should necessarily all be equal in respect of size, or some other kind of magnitude, is as irrelevant to the true nature of a fraction as the assumption of five things necessarily meaning five equal things is to the true nature of the number five.

Since $(a, 1)$ is characteristic of an aggregate of a things each of which is also regarded as a single object, it is clear that $(a, 1)$ is identical with a.

If we suppose each of the b elements in an aggregate, of which the cardinal number is b, to be replaced by an aggregate of n elements, we have now an aggregate with nb for its cardinal number, and instead of a elements chosen out of this aggregate we now have na of the new elements, each of which is to be regarded as associated with the cardinal number nb. We represent these na elements by (na, nb), which is equivalent to (a, b), since the two forms represent two different aspects of the same process. Therefore we have $(a, b) = (na, nb)$, or in the ordinary notation $a/b = na/nb$. This relation is in complete accordance with the law of logical (not arithmetical) addition, that a mere repetition of a term yields only the term itself.

Since $(a, b) = (ad, bd)$, and $(c, d) = (bc, bd)$, we regard (a, b) as greater, equal to, or less than (c, d), in the purely ordinal sense of the terms, according as ad is $\gtreqless bc$. For the two numbers (ad, bd), (bc, bd) are characteristic of the process of taking ad, bc elements respectively from an aggregate of the same cardinal number bd; and thus the relative order of the two numbers (a, b), (c, d) will naturally be fixed in accordance with the relative order of the two numbers ad, bc.

The addition of the two numbers (a, b) and (c, d) is equivalent to that of (ad, bd) and (bc, bd), and is consequently naturally defined as given by $(ad + bc, bd)$, which characterizes the amalgamation of two aggregates of which the numbers are ad, bc, the elements of each of which all belong to an aggregate of number bd, or to one of several such aggregates.

To interpret the operation of multiplication, let us consider an object represented by (c, d); this consists of c things each belonging to an aggregate of d things. To multiply it by (a, b), is to take a such objects each of which belongs to an aggregate of b such objects; we have on the whole one or more aggregates of bd elements, and out of these, ac elements are to be taken. Thus the multiplication of the number (c, d) by the number (a, b) may be understood to characterize the result of taking a objects, each of which is characterized by (c, d), out of one or more collections of b objects, each of which objects is characterized by (c, d). This is the same thing as the process of taking ac objects out of one or more aggregates of bd objects, and is characterized by the number (ac, bd); we are thus led to the law of multiplication

$$(a, b) \times (c, d) = (ac, bd), \text{ or } \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

NEGATIVE NUMBERS, AND THE NUMBER ZERO

14. Although the operation of addition is always possible within the aggregate of integral and fractional numbers, yet the inverse operation of subtraction is not always possible, thus a number x cannot be found such that $x + (c, d) = (a, b)$, unless $(a, b) > (c, d)$. As the limitation of the possibility of division suggested the introduction of fractional numbers, so this limitation of the possibility of subtraction suggests the introduction of a further set of new numbers, which shall be such that, within the so completed aggregate, subtraction may always be a possible operation.

If $\alpha, \beta, \gamma, \delta$ denote integral or fractional numbers such that $\alpha > \beta, \gamma > \delta$; we may put $\alpha = \beta + x$, $\gamma = \delta + y$; then $x = \alpha - \beta$, $y = \gamma - \delta$. We have

$$\alpha + \gamma = \beta + \delta + x + y,$$

hence
$$x + y = (\alpha + \gamma) - (\beta + \delta),$$

or
$$(\alpha - \beta) + (\gamma - \delta) = (\alpha + \gamma) - (\beta + \delta) \quad \dots\dots\dots\dots (1).$$

Again, if $\alpha - \beta = \gamma - \delta$, *i.e.* $x = y$, we have $\alpha + \delta = \beta + \delta + x = \beta + \gamma$,

or
$$\alpha + \delta = \beta + \gamma, \text{ if } \alpha - \beta = \gamma - \delta \quad \dots\dots\dots\dots (2).$$

Lastly, we have

$$\begin{aligned} \alpha\gamma = (\beta + x)(\delta + y) &= \beta(\delta + y) + x(\delta + y) \\ &= \beta\delta + \beta y + x\delta + xy; \end{aligned}$$

hence
$$\begin{aligned} \alpha\gamma + \beta\delta &= \beta(y + \delta) + \delta(x + \beta) + xy \\ &= \beta\gamma + \alpha\delta + xy; \end{aligned}$$

hence
$$(\alpha - \beta)(\gamma - \delta) = (\alpha\gamma + \beta\delta) - (\alpha\delta + \beta\gamma) \quad \dots\dots\dots\dots (3).$$

The rules (1), (2), (3), with regard to the numbers $\alpha - \beta$, $\gamma - \delta$, which so far exist only when $\alpha > \beta$, $\gamma > \delta$, suggest the mode of the extension referred to above.

15. Let α, β be any two numbers integral or fractional, and conceive a new number $D(\alpha, \beta)$, formed by the association of α and β, to be defined as subject to the laws

$$\begin{aligned} &(4),\ D(\alpha, \beta) = D(\gamma, \delta), \text{ if } \alpha + \delta = \beta + \gamma, \\ &(5),\ D(\alpha, \beta) + D(\gamma, \delta) = D(\alpha + \gamma, \beta + \delta), \\ &(6),\ D(\alpha, \beta) \times D(\gamma, \delta) = D(\alpha\gamma + \beta\delta, \alpha\delta + \beta\gamma): \end{aligned}$$

it will be observed that when $\alpha > \beta$, and $\gamma > \delta$, $D(\alpha, \beta)$ may denote $\alpha - \beta$, the three laws becoming (2), (1), (3). It will now be shewn that the symbol $D(\alpha, \beta)$ defines a number of an aggregate within which the operation of subtraction is always possible. For, to find a number x, such that

$$x + D(\alpha, \beta) = D(\gamma, \delta), \text{ we see that } x = D(\beta + \gamma, \alpha + \delta),$$

since
$$D(\alpha + \beta + \gamma, \alpha + \beta + \delta) = D(\gamma, \delta), \text{ in virtue of (4).}$$

Since $D(\alpha, \alpha) = D(\gamma, \gamma)$, we see that $D(\alpha, \alpha)$ is independent of α; and thus $D(\alpha, \alpha)$ defines a new number which is called the *number zero, and is denoted by the symbol* 0.

The number zero is regarded as characteristic of the absence of all elements from an aggregate of which the existence has been contemplated; it is the number of such a hypothetical aggregate, in a sense similar to that in which an integer is the number of an actual aggregate.

The number $D(\alpha + k, k)$ depends only on α, and we shall postulate that it is identical in meaning with α itself.

The numbers $D(\alpha, \beta)$, or $\alpha - \beta$, for which $\alpha > \beta$, are called positive numbers, and form the aggregate of integral and fractional numbers we have previously considered.

Those numbers, for which $\alpha < \beta$, are called negative numbers

Since by (5), $D(\alpha, \beta) + D(\beta, \alpha) = D(\alpha + \beta, \alpha + \beta) = 0$, the number $D(\beta, \alpha)$ may be denoted by $-D(\alpha, \beta)$, or in ordinary notation $-(\alpha - \beta)$. Thus *to every positive number x there corresponds a single negative number $-x$, which is such that $x + (-x) = 0$.*

We may now use the notation $\alpha - \beta$ in every case for $D(\alpha, \beta)$, and thus

$$\alpha - \beta = -(\beta - \alpha).$$

From (6), it is seen that the operation of division is always possible for two members of the complete aggregate of positive and negative numbers and zero, except when the divisor is the number zero, in which case the operation is meaningless.

From (6), we see by putting $\gamma = \delta$, $N = D(\alpha, \beta)$, that $N.0 = 0$. From (5), we have $N + 0 = N$.

Any number $D(\alpha, \beta)$ is said to be greater in the ordinal sense than $D(\gamma, \delta)$, when $D(\alpha, \beta) - D(\gamma, \delta)$ is positive; thus the complete aggregate of positive and negative integral and fractional numbers together with the number zero is one in which all the numbers are arranged in a definite order. This aggregate is known as the *aggregate of rational numbers.*

In the aggregate of rational numbers so ordered, the number zero has lower rank than any of the positive numbers, and higher rank than any of the negative numbers. Further, if x, y are two positive numbers of which x has higher rank than y, the negative number $-x$ has lower rank than $-y$.

If x, y are any two rational numbers, such that $x < y$, there exist an unlimited number of rational numbers each of which is $> x$, and $< y$.

Such numbers are said to be *between* x and y. For it can be seen at once, from the definition of order given above, that $\frac{1}{2}(x + y)$ is one such number; between $\frac{1}{2}(x + y)$ and either x or y another rational number can, in a similar manner, be found. This process can be carried on without end: and it is clear that, in accordance with the mode of ordering of the aggregate, defined above, all the numbers thus determined are between x and y.

If x, y are any two positive rational numbers such that $x < y$, an integer n can be found which is such that $nx > y$.

For $\frac{y}{x}$ is a rational number such that $\frac{y}{x}.x = y$, again if $\frac{p}{q}$ be any positive rational number, there exist integers which are $> \frac{p}{q}$; for $p + 1$ is itself such an integer. If n is an integer which is $> \frac{y}{x}$, we have $nx > \frac{y}{x}.x = y$: thus the theorem is established.

IRRATIONAL NUMBERS

16. The only numbers of which the existence was recognized by the Greek geometers were the rational numbers, although the fact that the ratio of two geometrical magnitudes is not necessarily exactly representable by such a number appears to have been discovered at a very early period. Euclid gave, in the fifth book of his treatise, a discussion of the theory of ratios, and in the tenth book, a theory of those incommensurable magnitudes which are ideally constructible by means of straight lines and circles. In later times*, the idea was current that, to the ratio of any two magnitudes of the same kind, there corresponds a definite number: and in fact Newton in his *Arithmetica Universalis* expressly defines a number as the ratio of any two quantities. Before the recent development of the arithmetical theories of irrational number, and to a considerable extent even later, a number has been regarded as the ratio of a segment of a straight line to a unit segment, and the conception of irrational number as the ratio of incommensurable segments has been accepted as a sufficient basis for the use of such numbers in Analysis.

In accordance with the doctrine that Mathematical Analysis must rest upon a purely arithmetical basis, the introduction of irrational numbers into Analysis must be made without an appeal to our intuition of extensive magnitude, but rather by an extension of the conception of Number, resting on a further development of the ideas which have been here discussed in connection with the theory of rational numbers. The necessity for this extension of the domain of Number arises not only on account of the inadequacy of rational numbers for application to ideally exact measurement, but also, as will be explained later in detail, because the theory of limits, which is an essential element in Analysis, is incapable of any rigorous formulation apart from a complete arithmetical theory of irrational numbers.

---

* A good short account of the history of this subject will be found in the article I. A. 3, "Irrationalzahlen und Konvergenz unendlicher Prozesse," by A. Pringsheim, in the *Encyclopadie der math. Wissenschaften*, vol. I. See also M. Cantor, *Geschichte der Math.*, vol. I.

Before the recent establishment of the theory of irrational numbers, no completely adequate theory of Magnitude was in existence. This is not surprising, if we recognize the fact that the language requisite for a complete description of relations of magnitudes must be provided by a developed Arithmetic.

17. The successive extensions of the domain of Number, by the introduction of fractional and of negative numbers, were suggested by the desirability of so completing the domain that the operations of division and subtraction, which are not always possible in the more limited domain, might always be so in the more extended one. In the aggregate of rational numbers, the operations of addition, subtraction, multiplication, and division (except when the divisor is zero) are always possible operations; but it can be readily shewn that the inverse operation involved in determining a fractional power of a rational number is not, in general, a possible one.

As the simplest case of this impossibility of such operation, we may take the problem of finding the square root of a positive integer m which is not a square number. It can be shewn that such a number has no square root within the aggregate of rational numbers.

If possible* let m be the square of a rational fraction p/q in its lowest terms; thus $p^2 - mq^2 = 0$. There always exists a positive integer λ such that $\lambda^2 < m < (\lambda + 1)^2$; we then have $\lambda q < p < (\lambda + 1)\, q$.

Now let us consider the identity

$$(mq - \lambda p)^2 - m\,(p - \lambda q)^2 = (\lambda^2 - m)\,(p^2 - mq^2) = 0$$

From this identity it follows that m is the square of the rational number $(mq - \lambda p)/(p - \lambda q)$, of which the denominator is less than q, and this is contrary to the hypothesis that m is the square of the fraction p/q which is in its lowest terms. It thus appears that there exists no rational number of which the square is m.

On the formal side of Arithmetic, a demand for the extension of the domain of number arises from the impossibility of carrying out, with the requisite generality, certain operations, as in the example given above. Such extensions of the domain of number as are made when fractional, negative, irrational, and complex numbers are successively adjoined to the original integral numbers, are made in accordance with a principle known as that of the permanence of forms, which was first indicated by Peacock†, and further developed by Hankel[*]. This principle may be stated in the form that, in order to generalize the conception of number, the following four requisites must be satisfied:

---

* This proof is given by Dedekind in his tract *Stetigkeit und irrationale Zahlen*. An extension of Dedekind's method to the case of nth roots has been given by S. M. Jacob; see *Proc. Lond. Math. Soc.* (2), vol. I (1903), p. 166.

† *British Association Report* for 1834; also *Symbolical Algebra*, Cambridge (1845).

(1) Every operation which is represented by a formal expression involving the unextended class of numbers, and which does not result in the representation of a number of the unextended class, must have a meaning assigned to it of such a character that the formal expression may be dealt with according to the same rules as would be applicable if the expression represented one of the unextended class of numbers.

(2) An extended definition of number must be given, such that a formal expression, as in (1), may represent a number in the extended sense of the term.

(3) A proof must be given that for numbers of the extended class the same formal laws of operation hold as for the unextended class.

(4) Definitions must be given of the meaning of greater, equal, and less, in the extended domain of number, these terms being taken in the ordinal sense.

The arithmetical theory of irrational numbers has been developed in three main forms, of which the first[†] was given by Weierstrass in his lectures on Analytical Functions; the second[‡] is that of G. Cantor, which was developed in further detail by Heine[§], and was also developed independently by Ch. Méray[||]; the third, that of R. Dedekind[¶], appeared about the same time as that of Cantor. We shall give an account of the theories of Dedekind and of Cantor, and shall shew that they are fundamentally identical.

KRONECKER'S SCHEME OF ARITHMETIZATION

18. As it is now generally understood, the term "arithmetization" is used to denote the movement which has resulted in placing analysis on a basis free from all notions derived from the idea of measurable quantity, the fractional, negative, and irrational numbers being so defined that they depend ultimately upon the conception of integral number. An extreme theory of arithmetization has however been advocated by Kronecker[**], who proposed the abolition of all modifications and extensions of the conception of number, the integral numbers being alone retained. His ideal* is that every theorem in analysis shall be stated as a relation between integral numbers only, the terminology involved in the use of negative, fractional, and irrational numbers, being entirely removed. This ideal, if it were possible to attain it, would amount to a reversal of the actual historical course which the science has pursued, for all actual progress has depended upon successive generalizations of the notion of number, although these generalizations are now regarded as ultimately dependent on the whole number for their foundation. The abandonment of the inestimable advantages of the formal use in Analysis of the extensions of the notion of number could only be characterized as a species of Mathematical Nihilism.

---

[*] See his *Theorie der komplexen Zahlsysteme*, Leipzig (1867).

[†] For an account of this Theory see S. Pincherle, *Giorn. di mat.*, vol. xviii (1880), p. 185; also O. Biermann, *Theorie der analytischen Funktionen*, Leipzig (1887), p. 19.

[‡] *Math. Annalen*, vol. v (1872), p. 123; see also *Math. Annalen*, vol. xxi (1883), p. 545, where Cantor discusses all the three theories.

[§] *Crelle's Journal*, vol. lxxiv (1872), p. 172.

[||] *Nouveau Précis d'Analyse infinitésimale*, Paris (1872).

[¶] *Stetigkeit und irrationale Zahlen*, Brunswick (1872).

[**] See *Crelle's Journal*, vol. ci (1887), p. 337, "Ueber den Zahlbegriff."

THE DEDEKIND THEORY OF IRRATIONAL NUMBERS

19. Let us consider the aggregate of all the rational numbers ordered in the manner which has been previously discussed, and let us take any one such number N. We may conceive all the rational numbers to be divided into two classes R_1, and R_2, such that every number of R_1 is, in the ordinal sense of the term, less than every number belonging to the second class R_2, the two classes being separated by the number N, which may itself be assigned at choice either to the first or to the second class. If N belongs to the first class, it is the greatest number in that class, and the numbers of the second class have no number which is less than all the others of that class; if N be taken to belong to the second class, it is the least number in that class, and there exists no number in the first class which is greater than all the others; for if any rational number less than N be taken, it is always possible to find another greater one which is less than N. Such a division of the rational numbers into two classes is called a *section* (Schnitt), and we therefore say that, *corresponding to any given rational number there exists a section which divides the aggregate of rational numbers into two classes, such that all the numbers of the first class are less than all those of the second class; and such that either in the first class there is no greatest number, or else in the second class there is no least number.*

It can be shewn, by means of examples, that sections of the aggregate of rational numbers exist which are different in character from those just described. If m is a positive integer which is not a square number, we may conceive the rational numbers to be divided into two classes, the first of which contains all the negative numbers and also those positive numbers of which the square is less than m, including zero; the second class contains all the positive numbers of which the square is greater than m. The first class contains no greatest number, and the second class contains no least number; this section is said to be related to an irrational number $\sqrt{m}$, in the same way as a section such as has been considered above is related to a rational number. This example shews that sections of the rational numbers R exist, such that R is divided into two classes R_1, R_2, where every number of R_1 is less than every number of R_2, and such that R_1 contains no number greater than all the others, and also R_2 contains no number less than all the others.

---

* He writes (*loc. cit.* p. 338)· "Und ich glaube auch, dass es dereinst gelingen wird, den gesammten Inhalt aller dieser mathematischen Disciplinen zu 'arithmetisiren,' d. h. einzig und allein auf den im engsten Sinne genommenen Zahlbegriff zu gründen, also die Modificationen und Erweiterungen dieses Begriffs (ich meine hier namentlich die Hinzunahme der irrationalen sowie der continuirlichen Grossen) wieder abzustreifen, welche zumeist durch die Anwendungen auf die Geometrie und Mechanik veranlasst worden sind." He proceeds to shew in detail, how the notions of negative, fractional, and algebraical numbers can be avoided by substituting for equalities in which these numbers occur, congruences relative to certain moduli or systems of moduli. A similar suggestion has been made by Cauchy with reference to imaginary numbers.

A new aggregate of objects, the *real numbers*, may now be defined as follows:

To every section (R_1, R_2) *of the aggregate* R *of rational numbers, such that every number of* R *belongs to one or other of the two classes* R_1, R_2, *and every number in* R_1 *is ordinally less than every number in* R_2, *there corresponds a real number.*

In case neither R_1 *contains a number which is ordinally greater than all the others in* R_1, *nor* R_2 *contains a number which is ordinally less than all the others in* R_2, *the real number corresponding to the section is said to be an irrational number.*

In case either R_1 *has a greatest number* x, *or* R_2 *has a least number* x, *the section is said to define a real number corresponding to the rational number* x.

The *real* number which corresponds to a *rational* number x, though conceptually distinct from x, has no properties distinct from those of x, and is usually denoted by the same symbol.

The definition of a real number can be put into a different and somewhat less abstract form, by employing the notion of a lower segment of the aggregate of rational numbers. A *lower segment of the aggregate* R *of rational numbers* is any class of rational numbers which contains no number greater than all the others, and such that if any number whatever of the class be taken, the class contains *all those numbers of* R which are less than that number. A lower segment of R is identical with one of Dedekind's classes R_1, in case R_1 contains no greatest number.

A real number may be defined to be a lower segment of the aggregate* R *of rational numbers; and thus every real number, whether irrational or not, is a definite class of rational numbers.*

---

* This form of the definition is that given by B. Russell, see *The Principles of Mathematics*, vol. I (1903), chaps XXIII and XXIV; it was suggested by Peano, see *Rivista di Matematica*, vol. VI (1896–99), pp. 126–140.

In accordance with this definition, the *real* number 3, for example, is defined to be the aggregate of all rational numbers which are less than the *rational* number 3; the irrational number $\sqrt{3}$ is defined as the aggregate of all rational numbers which are either negative, or if positive have their squares less than 3, the number zero being also included in the aggregate.

The use, here adopted, of the term *real* number, is sanctioned by general usage. The employment of the term *real* has originated from the contrasting of these numbers, not with rational numbers, but with complex numbers. The extension of the term Number to the real numbers is justified by the fact that it is possible to define the operations of addition, multiplication, &c., for real numbers, so that the formal laws of these operations are in agreement with those which hold for operations within the domain of the rational numbers.

20. It will now be shewn that the aggregate of real numbers, defined in Dedekind's manner, can be so ordered that every real number has a definite rank in the aggregate, *i e.* of any two real numbers it is determinate which has the higher and which the lower rank.

The basis of the scheme of order being taken to be the ordered aggregate of rational numbers, let us denote by n, n' any two real numbers, and let the sections by which they are defined be denoted by (R_1, R_2), (R_1', R_2') respectively.

The following cases may arise:

(1) If (R_1, R_2) and (R_1', R_2') are identical, that is, if every number in R_1 is also in R_1', and every number in R_2 is also in R_2', the two numbers n, n' are identical; thus $n \equiv n'$.

(2) Let us next suppose that there is one rational number $r_1 \equiv r_2'$, which is contained in R_1, but not in R_1'; it is consequently contained in R_2'. All the numbers in R_1' are less than r_2', and hence all the numbers in R_1' are in R_1. Since r_1 is the only number in R_1 which is contained in R_2', it follows that r_1 is greater than all the other numbers in R_1, and thus the number n defined by (R_1, R_2) is a number corresponding to the rational number r_1 or r_2'. All the elements in R_1' are contained in R_1, and are less than r_2'; all the numbers in R_2' except r_2', are greater than r_2', for if not they would be contained in R_1: hence the section (R_1', R_2') defines the real number $n' \equiv n$, corresponding to the rational number $r_2' \equiv r_1$. The two sections are essentially identical, the only difference being that the rational number $r_1 \equiv r_2'$, is regarded as belonging to the first class in one section and to the second class in the other section.

(3) If there are two different numbers belonging to R_1 which also belong to R_2', there are an indefinite number of other numbers which have the same property, since an unlimited number of rational numbers can be

found which lie between two given rational numbers. In this case we define the real number n or (R_1, R_2) to be greater, in the ordinal sense of the term, than n' or (R_1', R_2'), agreeably with the definition already given for the rational numbers.

The cases in which one, or more than one, number which belongs to R_1' also belongs to R_2, may be treated in a similar manner, thus we define the meaning of the relation $n < n'$. It is easily seen that if $n > n'$, and $n' > n''$, then the relation $n > n''$ is also satisfied. Thus the system of real numbers is arranged in a regular order, such that those of them which correspond to rational numbers have the same relative rank as the corresponding rational numbers have in the aggregate of rational numbers.

21. The aggregate of real numbers has the following properties:

(1) If $\alpha > \beta$, and $\beta > \gamma$, then $\alpha > \gamma$.

(2) Between any two real numbers α, γ there are an unlimited number of real numbers. This is easily proved from the corresponding property of rational numbers, by considering the sections which define the numbers.

(3) If α is a fixed real number, then all real numbers may be divided into two classes $\bar{R}_1$, $\bar{R}_2$, such that $\bar{R}_1$ contains all the real numbers which are less than α, and $\bar{R}_2$ contains all the real numbers which are greater than α. The number α may be regarded either as belonging to $\bar{R}_1$, in which case it is the greatest number in $\bar{R}_1$, or else as belonging to $\bar{R}_2$, in which case it is the least number in $\bar{R}_2$. This also follows from the definition above.

(4) If the aggregate of real numbers falls into two classes $\bar{R}_1$, $\bar{R}_2$, such that every number of $\bar{R}_1$ is less than every number of $\bar{R}_2$, then there exists one, and only one, number by which this section is produced.

To prove this, we observe that the section $(\bar{R}_1, \bar{R}_2)$ of the aggregate of real numbers also defines a section (R_1, R_2) of the aggregate of rational numbers, such that all rational numbers belonging to R_1 correspond to real numbers which belong to $\bar{R}_1$, and all numbers belonging to R_2 correspond to real numbers which belong to $\bar{R}_2$.

Let N be the real number defined by the section (R_1, R_2), and let N' be any real number different from N, defined by the section (R_1', R_2'). There are an indefinite number of rational numbers n which belong to only one of the aggregates R_1, R_1'; let $\bar{n}$ be the real number corresponding to n. If $N' < N$, then n belongs to R_1, and therefore $\bar{n}$ belongs to $\bar{R}_1$; and since $N' < \bar{n}$, it follows that N' belongs to $\bar{R}_1$. Similarly, if $N' > N$, we can shew that N' belongs to $\bar{R}_2$. It has thus been shewn that every number different from N belongs to $\bar{R}_1$ or to $\bar{R}_2$, according as it is less or greater than N. Thus N is either the greatest number in $\bar{R}_1$ or the least in $\bar{R}_2$, and therefore N is the only number by which the section $(\bar{R}_1, \bar{R}_2)$ can be made.

22. The operations between two real numbers may, in accordance with the above definition of real numbers by means of sections, be so defined that the result of each operation corresponds to a section of the rational numbers: thus the arithmetical operations are reduced to operations with rational numbers.

A complete theory of the operations involving real numbers can be established; and the formal laws of the operations can be shewn to be the same as in the case of the rational numbers, the range of possibility of operations being greater in the case of real than in that of rational numbers. This theory has been worked out to some extent by Dedekind: but as the Cantor theory of real numbers lends itself to a simpler detailed treatment of the operations than that of Dedekind, and as it will appear that the two theories are fundamentally equivalent to one another, it will be sufficient, as an example of the general method of treating operations in accordance with Dedekind's theory, to take only the case of the addition of two real numbers.

Let a, b be two real numbers defined by means of the sections $(R_1 . R_2)$. (R_1', R_2') respectively; then the sum $a + b$, of a and b, is defined by means of a section (R_1'', R_2'') which satisfies the following conditions:—If c_1 is any rational number, it is put into the class R_1'', provided there are two rational numbers a_1 in R_1, and b_1 in R_1', such that $a_1 + b_1 \geqq c_1$; all rational numbers c_2 for which this is not the case fall into the class R_2''. It is clear that every number c_1 is less than every number c_2, hence the section (R_1'', R_2'') is defined by means of this condition.

It can be shewn that, when a, b both correspond to rational numbers, this definition is in agreement with the ordinary definition of the sum of two rational numbers, so that the sum of the numbers corresponds to the sum of the corresponding rational numbers. Every number c_1 in R_1'', is $\leq a + b$, because $a_1 \leq a$, $b_1 \leq b$, and therefore $a_1 + b_1 \leq a + b$. Further, if there were contained in R_2'' a number $c_2 < a + b$, so that $a + b = c_2 + p$, where p is a positive rational number, we should have

$$c_2 = (a - \tfrac{1}{2}p) + (b - \tfrac{1}{2}p),$$

and this is contrary to the definition of c_2, because $a - \frac{1}{2}p$ belongs to R_1, and $b - \frac{1}{2}p$ to R_1'; thus every number c_2 in R_2'', is $\geq a + b$, and it has consequently been shewn that (R_1'', R_2'') defines the number $a + b$. As is usual, we have denoted the rational numbers a, b and the conceptually distinct real numbers a, b by the same symbols.

THE CANTOR THEORY OF IRRATIONAL NUMBERS

23. The Cantor theory of irrational numbers essentially depends upon the use of convergent simply infinite ascending aggregates, or convergent sequences (Fundamentalreihen) in which the elements are rational numbers; we therefore proceed to define and discuss these aggregates.

A simply infinite ascending aggregate $(a_1, a_2, a_3, \ldots a_n, \ldots)$, *in which each element is a rational number, is said to be convergent, if it is such that, corresponding to any fixed arbitrarily chosen positive rational number* ϵ, *as small, in the ordinal sense, as we please, an integer* n *can be found such that* $| a_n - a_{n+m} | < \epsilon$, *for* $m = 1, 2, 3, \ldots$.

The symbol $| x |$ is here used to denote that one of the two numbers x, $-x$ which is positive; $| x |$ is said to be the absolute value of x.

This definition is equivalent to the statement that, in a simply infinite convergent aggregate, an element can always be found whose absolute difference from any element whatever which comes after it is as small as we please.

It should be observed that the terms "as small as we please," or "arbitrarily small," as applied to a positive number which is at choice, have reference to the conception of order only, and not to the non-arithmetical notion of magnitude. These expressions denote only that the number can be so chosen as to be of lower rank than any other arbitrarily chosen positive number.

To each value of ϵ there corresponds a value of n, which will in general have to be increased when ϵ is made smaller.

We may denote the aggregate by the symbol $\{a_n\}$, and shall speak of it shortly as a convergent sequence; that it is simply infinite will in future be understood.

For a convergent sequence, corresponding to any arbitrarily chosen positive number ϵ, *an integer* n *can be found, such that, from and after that value of* n, *the absolute difference of any two elements is less than* ϵ.

For choose n so that $| a_n - a_{n+m} | < \frac{1}{2}\epsilon$, for all positive integral values of m; then $| a_{n+m} - a_{n+m'} | \leqq | a_n - a_{n+m} | + | a_n - a_{n+m'} | < \epsilon$.

For the convergent sequence $\{a_n\}$, if we choose n such that

$$| a_n - a_{n+m} | < \epsilon,$$

then for $m = 1, 2, 3, \ldots$, the value of a_{n+m} for all values of m, lies between $a_n + \epsilon$, and $a_n - \epsilon$; that is to say, from and after some value of n, all the elements lie between two rational numbers whose difference is arbitrarily small. There exist therefore two positive numbers α, α', of which the smaller α' may be zero, such that, from and after some fixed value of n, say n_1, all the elements lie in absolute value between α and α'.

It follows that:

If $\{a_n\}$ *be a convergent sequence, there exists a positive number* N *such that* $| a_n | < N$, *for every value of* N.

For N need only be taken to be greater than all the numbers

$$| a_1 |, | a_2 |, \dots | a_{n_1} |, a.$$

24. *If the aggregate* $(a_1, a_2, \dots a_n, \dots)$ *is such that, from and after some fixed element, each element is not greater than the following one, and if all the elements are less than some fixed number* N, *then the aggregate is a convergent sequence.*

For if the aggregate is not convergent, there must exist some positive number δ, such that an indefinite number of increasing values $n_0, n_1, n_2, \dots$ of n can be found, for which $| a_{n_1} - a_{n_0} |$, $| a_{n_2} - a_{n_1} |$, $| a_{n_3} - a_{n_2} | \dots$ are all $> \delta$. Since $a_{n_1} - a_{n_0}, a_{n_2} - a_{n_1}, \dots$ are all positive, we have $a_{n_r} \geq a_{n_0} + r\delta$, where r can always be taken so large that $a_{n_0} + r\delta > N$, or $a_{n_r} > N$, which is contrary to the hypothesis. Hence the aggregate is convergent.

It may in a similar manner be shewn that the aggregate is convergent if, from and after some fixed element, each element is not less than the following one, and if all the elements are greater than some fixed number.

An aggregate $(a_1, a_2, \dots a_n, \dots)$ such that $a_1 \leq a_2 \leq a_3 \leq \dots$ is said to be a *monotone non-diminishing sequence*. If all the elements are less than some fixed number N, the sequence is convergent. Similarly if

$$a_1 \geq a_2 \geq a_3 \geq \dots$$

the aggregate is said to be a *monotone non-increasing sequence*. It is convergent if all the numbers a_n are greater than some fixed number.

If $\{a_n\}$, $\{b_n\}$ *are two convergent sequences of rational numbers, a value of* n *can be found corresponding to any arbitrarily assigned number* ϵ, *such that both* $| a_{n+m} - a_{n+m'} |$ *and* $| b_{n+m} - b_{n+m'} |$ *are less than* ϵ, m *and* m' *having all positive values.*

For we have only to choose for n the greater of the two values corresponding to ϵ, for each aggregate separately.

25. It will now be shewn that *the aggregates*

$$\{a_n + b_n\}, \{a_n - b_n\}, \{a_n b_n\}, \left\{\frac{a_n}{b_n}\right\},$$

in which the elements are the sum, difference, product, and quotient, respectively of the corresponding elements of the two convergent sequences $\{a_n\}$, $\{b_n\}$, *are also convergent sequences; with a certain restriction in the last case.*

We have $| (a_n \pm b_n) - (a_{n+m} \pm b_{n+m}) | \leq | a_n - a_{n+m} | + | b_n - b_{n+m} |$; now n can be so chosen for a given ϵ, that for all values of m,

$$| a_n - a_{n+m} | < \tfrac{1}{2}\epsilon,$$

and $|b_n - b_{n+m}| < \frac{1}{2}\epsilon$, hence so that $|(a_n \pm b_n) - (a_{n+m} \pm b_{n+m})| < \epsilon$: therefore the aggregates $\{a_n + b_n\}$, $\{a_n - b_n\}$ are convergent.

Again,

$$|a_nb_n - a_{n+m}b_{n+m}| = |a_n(b_n - b_{n+m}) + b_{n+m}(a_n - a_{n+m})|$$
$$< \alpha|b_n - b_{n+m}| + \beta|a_n - a_{n+m}|,$$

where α, β are the two positive numbers which are such that $|a_n| < \alpha$, $|b_{n+m}| < \beta$, for all values of n and m.

We can take n so large that $|b_n - b_{n+m}| < \delta$, $|a_n - a_{n+m}| < \delta$, where δ is at our choice, and may be taken to be $\frac{\epsilon}{\alpha + \beta}$. Hence, for this value of n, $|a_nb_n - a_{n+m}b_{n+m}| < \epsilon$, for every value of m: and thus $\{a_nb_n\}$ has been shewn to be a convergent sequence.

Lastly, in the case of $\left\{\frac{a_n}{b_n}\right\}$, we shall suppose that all the elements of $\{b_n\}$ are numerically greater than some fixed positive number β'.

We have then

$$\frac{a_n}{b_n} - \frac{a_{n+m}}{b_{n+m}} = \frac{a_n(b_{n+m} - b_n) + b_n(a_n - a_{n+m})}{b_nb_{n+m}},$$

hence $$\left|\frac{a_n}{b_n} - \frac{a_{n+m}}{b_{n+m}}\right| < \frac{\alpha|b_n - b_{n+m}| + \beta|a_n - a_{n+m}|}{\beta'^2}.$$

If now n be chosen so that $|b_n - b_{n+m}|$, $|a_n - a_{n+m}|$ are both less than $\frac{\beta'^2}{\alpha + \beta}\epsilon$, for every value of m, then for such a value of n,

$$\left|\frac{a_n}{b_n} - \frac{a_{n+m}}{b_{n+m}}\right| < \epsilon;$$

therefore $\left\{\frac{a_n}{b_n}\right\}$ is a convergent sequence, provided $|b_n|$ is, for all values of n, greater than some fixed positive number β', which may be as small as we please, but must not be zero.

26. The essence of Cantor's theory consists in the postulation of the existence of an aggregate of objects for thought, the real numbers, ordered in a definite manner, which manner is assigned by means of certain prescribed rules. Any element of the aggregate of real numbers is regarded as capable of symbolical representation by means of a convergent sequence of which the elements are rational numbers; and the mode in which the aggregate of real numbers is ordered is specified by means of formal rules relating to these convergent sequences. The aggregate of real numbers contains within itself an aggregate of objects which is similar to the ordered aggregate of rational numbers which has already been considered, in the sense that to each rational number there corresponds a certain real number; and the relative order of any two rational numbers, in the ordered aggregate of rational numbers, is the same as the relative order of the two corresponding real numbers in the new aggregate of real numbers. The rational numbers are frequently regarded as identical with the real numbers to which they correspond, and are denoted by the same symbols. In the development of Analysis, this identity leads to no difficulties; but, in the fundamental theory of the aggregate of real numbers, a conceptual distinction between rational numbers and the real numbers to which they correspond must be made, in order to obviate logical difficulties, and especially with a view to coordinating Cantor's theory with that of Dedekind. Those real numbers which do not correspond to rational numbers are called irrational numbers; and those real numbers which correspond to rational numbers are usually spoken of as themselves rational numbers.

The rules by which the order of the real numbers in their aggregate is assigned are the following:

(1) Any convergent sequence $\{a_n\}$, of which the elements are rational numbers, is taken to represent a real number, which we may denote by a. Two such aggregates $\{a_n\}$, $\{b_n\}$ are taken to represent the same real number provided they satisfy the condition that, for any arbitrarily chosen positive rational number ϵ, a value of n can be found such that $| a_{n+m} - b_{n+m} | < \epsilon$, for this value of n, and for all values 0, 1, 2, 3, ..., of m. Symbolically*, we have $\{a_n\} \equiv \{b_n\}$ under the condition stated.

(2) The real number represented by $\{a_n\}$ is regarded as of higher rank, or in the ordinal sense greater, than the real number represented by $\{b_n\}$, if a value of n can be found such that $a_{n+m} - b_{n+m}$ is, for this value of n, and for all values 0, 1, 2, 3, ..., of m, greater than some fixed positive rational number δ. If n can be so determined that $a_{n+m} - b_{n+m}$ is negative and numerically greater than some fixed positive rational number δ, for every value of m, the number represented by $\{a_n\}$ is taken to be less than that represented by $\{b_n\}$.

The aggregate $(x, x, x, \ldots)$, or $\{x\}$, in which all the elements are identical with one rational number x, represents, since it is a convergent sequence, a real number which corresponds to the rational number x. It is clear, from the definition of order in (2), that the relative order of any two rational numbers, in the aggregate of rational numbers, is the same as that of the real numbers which correspond to them, in the aggregate of real numbers. The aggregate of rational numbers, and that of the real numbers which correspond to them, are *similar* aggregates.

Cantor's theory of irrational numbers, in the form in which it was presented by himself and by Heine, has been criticized† on the ground that an assumption is made that the sequence $\{x\}$, in which all the elements are the same rational number x, represents the rational number x itself, and that this amounts to an assumption that x is the limit of the sequence $\{x\}$; whereas the theory of arithmetical limits is represented by Cantor* as deducible from his theory of irrational numbers, and as not assumed in the construction of the theory itself. The theory in the form presented above is not open to this objection.

---

* Those who hold the view, advocated by Heine and others (see § 7, note), that a real number is identical with the set of symbols by which it is represented, can attach no direct meaning to this equality. It can only be taken to indicate that the two expressions may be used indifferently in any operation which involves the number.

† See B. Russell, *The Principles of Mathematics*, vol. I (1903), p. 285.

It can be shewn that any two convergent sequences $\{a_n\}$, $\{b_n\}$ satisfy one or other of the conditions laid down in the above definitions of equality and inequality, *i.e.* symbolically $\{a_n\} \gtreqless \{b_n\}$.

For, as has been shewn in § 23, corresponding to any arbitrarily chosen positive rational number δ, a value of n can be found such that a_{n+m} lies between $a_n + \delta$, and $a_n - \delta$, and such that, for the same value of n, b_{n+m} lies between $b_n + \delta$, and $b_n - \delta$; from this it follows that, for such value of n, $a_{n+m} - b_{n+m}$ lies between $a_n - b_n + 2\delta$ and $a_n - b_n - 2\delta$: or $a_{n+m} - b_{n+m}$ differs from $a_n - b_n$ by not more than 2δ. If corresponding values of δ and n can be found, for which $a_n - b_n + 2\delta$, $a_n - b_n - 2\delta$ have the same sign, then $a_{n+m} - b_{n+m}$ has the same sign as $a_n - b_n$, and is numerically greater than a fixed number; the condition of inequality of $\{a_n\}$, $\{b_n\}$ is then satisfied. If no such values of δ and n can be found, then $a_{n+m} - b_{n+m}$ is numerically less than 4δ; and since δ is arbitrarily small, the condition of equality of $\{a_n\}$, $\{b_n\}$ is then satisfied.

Although Cantor's form of the theory of irrational numbers, or rather of real numbers, is more convenient for detailed development than is Dedekind's form, yet it lies under the disadvantage that the nature of any single real number is veiled by the fact that, although it is a unique object, it is capable of representation by an unlimited number of convergent sequences, and therefore that the formal character of the theory does not make it clear what such a number really is. The comparison between the two theories which† will be given later on will throw light upon this point: for it will be shewn that a convergent sequence of the rational numbers is sufficient to define a section, of the kind fundamental in Dedekind's theory; and this, as we have seen, is equivalent to the definition of a lower segment, which is itself a certain definite class of rational numbers.

27. *The sum $a + b$, of two real numbers represented by the sequences $\{a_n\}$, $\{b_n\}$, is defined to be the real number represented by the sequence $\{a_n + b_n\}$; and the difference $a - b$ is defined as a number represented by $\{a_n - b_n\}$.*

It has been shewn in § 25 that the two sequences $\{a_n + b_n\}$, $\{a_n - b_n\}$ are convergent.

---

[*] See *Math. Annalen*, vol. XXI (1883), p. 568.

[†] In Tannery's work *Introduction à la théorie des fonctions d'une variable*, Paris (1886 and 1904), chap. I, the theory of irrationals is treated by a combination of the two methods of Cantor and Dedekind.

If $\{a_n\}$, $\{b_n\}$ represent the same number, the sequence $\{a_n - b_n\}$ defines the real number zero; for the condition that $|a_{n+m} - b_{n+m}| < \epsilon$, where ϵ is arbitrarily small, for a sufficiently great value of n, and for $m = 0, 1, 2, 3, \ldots$, is in this case satisfied.

The product ab, of two real numbers, is defined to be the number represented by the sequence $\{a_n b_n\}$, which has been shewn in § 25 to be convergent.

The quotient a/b is defined to be the number represented by the convergent sequence $\left\{\frac{a_n}{b_n}\right\}$.

The only restriction on this last definition is that b is not to be zero; for, when this condition is satisfied, the elements of the sequence $\{b_n\}$, which represents b, can be so chosen as to satisfy the restrictive condition given in § 25, that $\left\{\frac{a_n}{b_n}\right\}$ may be convergent.

It is necessary to shew that the sum $a + b$, the difference $a - b$, the product ab, and the quotient $\frac{a}{b}$, of two numbers a, b, as they have been defined above, are definite numbers independent of the particular convergent sequences used to represent the numbers a and b. Thus it must be shewn that if $\{a_n\} = \{a_n'\}$, $\{b_n\} = \{b_n'\}$, then

$$\{a_n + b_n\} = \{a_n' + b_n'\},\ \{a_n - b_n\} = \{a_n' - b_n'\},\ \{a_n b_n\} = \{a_n' b_n'\},$$

and

$$\left\{\frac{a_n}{b_n}\right\} = \left\{\frac{a_n'}{b_n'}\right\}.$$

We have

$$|(a_{n+m} \pm b_{n+m}) - (a'_{n+m} \pm b'_{n+m})| \leqq |a_{n+m} - a'_{n+m}| + |b_{n+m} - b'_{n+m}|.$$

Now n can be so chosen, corresponding to a fixed number ϵ, that

$$|a_{n+m} - a'_{n+m}| < \tfrac{1}{2}\epsilon,\ |b_{n+m} - b'_{n+m}| < \tfrac{1}{2}\epsilon, \text{ for } m = 0, 1, 2, 3, \ldots;$$

with this value of n, we now have the condition

$$|(a_{n+m} \pm b_{n+m}) - (a'_{n+m} \pm b'_{n+m})| < \epsilon,$$

satisfied; and this is the condition that $\{a_n \pm b_n\}$ represents the same number as $\{a_n' \pm b_n'\}$.

Again,

$$|a_{n+m}b_{n+m} - a'_{n+m}b'_{n+m}| \leqq |a_{n+m}(b_{n+m} - b'_{n+m})| + |b'_{n+m}(a_{n+m} - a'_{n+m})|$$
$$< A\,|b_{n+m} - b'_{n+m}| + B\,|a_{n+m} - a'_{n+m}|,$$

where A, B are fixed positive numbers. It is now clear that n may be so chosen that $|a_{n+m}b_{n+m} - a'_{n+m}b'_{n+m}| < \eta$, where η is an arbitrarily chosen positive number; thus $\{a_n b_n\}$, $\{a_n' b_n'\}$ represent the same number.

Again,

$$\left|\frac{a_{n+m}}{b_{n+m}} - \frac{a'_{n+m}}{b'_{n+m}}\right| \leqq \left|\frac{a_{n+m}}{b_{n+m}b'_{n+m}}(b_{n+m} - b'_{n+m})\right| + \left|\frac{b_{n+m}}{b_{n+m}b'_{n+m}}(a_{n+m} - a'_{n+m})\right|,$$

whence it can easily be seen that the condition is satisfied that

$$\begin{Bmatrix} a_{n+m} \\ b_{n+m} \end{Bmatrix} \text{ and } \begin{Bmatrix} a'_{n+m} \\ b'_{n+m} \end{Bmatrix}$$

represent the same number.

It is readily seen that the same commutative, associative, and distributive laws hold for the operations between real numbers as for those involving rational numbers.

28. *If, from and after some fixed element* a_n, *all the elements of* $\{a_n\}$ *are greater than some fixed positive rational number* δ, *then the real number represented by* $\{a_n\}$ *is positive, i.e. it is ordinally greater than zero.*

For, if we take any convergent sequence $\{b_n\}$ which defines the number zero, we have $\{a_n\} > \{b_n\}$; because, for some fixed value of n, $a_{n+m} - b_{n+m}$ is certainly positive for all values of m, and is greater than a fixed positive number; since n can be taken so large that $a_{n+m} > \delta$, and $b_{n+m} < \delta'$, where δ' is a positive rational number chosen less than δ.

Similarly, it may be shewn that *the number defined by* $\{a_n\}$ *is negative, if, from and after some fixed value of* n, *all the* a_n *are negative and numerically greater than some fixed positive rational number* δ.

The term "numerically greater" denotes that $|a_n| > \delta$ and thus refers to the absolute values of the numbers concerned.

It is easily seen that, *unless* $\{a_n\}$ *is such that, from and after some fixed value of* n, *all the elements have the same sign, then* $\{a_n\}$ *must represent the number zero.*

If $\{a_n\}$, $\{b_n\}$ *define two different real numbers* a, b, *then there lie between* a, b *an unlimited number of those real numbers which correspond to rational numbers.*

Suppose $a > b$, then there exist a definite rational positive number δ, and a least integer n, such that, for all positive integral values of m including zero, $a_{n+m} - b_{n+m} \geqq \delta$, $|a_n - a_{n+m}| < \epsilon$, $|b_n - b_{n+m}| < \epsilon$, where ϵ is a rational number chosen to be $< \frac{1}{2}\delta$. If we take any rational number x, which is $< \delta$, and $> \epsilon$, the number $\{a_n - x\}$, in which all the elements are identical, lies between $\{a_n\}$ and $\{b_n\}$, since, for every value of m, we have $a_{n+m} - (a_n - x) > x - \epsilon$; therefore a is greater than the real number which corresponds to $a_n - x$. Again,

$$(a_n - x) - b_{n+m} = (a_n - b_n) + (b_n - b_{n+m}) - x > \delta - \epsilon - x;$$

therefore, provided x is chosen to be $< \delta - \epsilon$, the real number which corresponds to $a_n - x$ is greater than b, and thus lies between a and b. The rational number $a_n - x$ may be chosen in an unlimited number of ways, since x is any rational number whatever which lies between $\delta - \epsilon$ and ϵ. To obtain one such number we may take $\epsilon = \frac{1}{4}\delta$, $x = \frac{1}{2}\delta$.

CONVERGENT SEQUENCES OF REAL NUMBERS

29. Convergent sequences will now be considered, of which the elements are real numbers. It might at first sight be imagined that we should be led, by the employment of such sequences, to a further extension of the domain of number; it will however be seen that this is not the case.

The definition of a convergent sequence of real numbers is precisely similar to the definition which has been given in the case of sequences of rational numbers: thus $(\alpha_1, \alpha_2, \dots \alpha_n, \dots)$ is a convergent sequence of real numbers, provided that, corresponding to each arbitrarily chosen positive real number η, a value of n can be found such that $|\alpha_n - \alpha_{n+m}| < \eta$, for $m = 1, 2, 3, \dots$. If we conceive that each such convergent sequence of real numbers represents a single ideal object, and if we give definitions of equality and inequality, and of the fundamental operations, precisely analogous to those given in § 26 and § 27, and assume as before that a convergent sequence in which all the elements are identical with the real number α is taken to represent that one of the new aggregate of objects which corresponds to α, it will be shewn that the new aggregate of objects is similar to the aggregate of real numbers, *i.e.* to each of the new objects there corresponds one of the real numbers, and also that the relation of order between corresponding pairs of elements in the two aggregates is the same It thus appears that the aggregate of new objects is practically identical with the aggregate of real numbers, since the two are ordinally similar. No such relation has been shewn to exist between the aggregate of real numbers and that of rational numbers; and it will be shewn later, in connection with the general theory of order-types, that no such relation of similarity can exist. Therefore the passage from rational numbers to real numbers involves a real extension of the domain of number; but the passage from real numbers to an aggregate of objects represented, in accordance with the rules referred to above, by convergent sequences of real numbers, does not lead to any essential extension of the domain of number.

Let $\{a_n\}$ be any convergent sequence of rational numbers, and let $\{\bar{a}_n\}$ denote the sequence of those real numbers which correspond to the rational numbers which form the elements of $\{a_n\}$. It can easily be shewn that $\{\bar{a}_n\}$ is a convergent sequence: for, if ϵ is an arbitrarily chosen positive rational number, and $\bar{\epsilon}$ the corresponding real number, the condition of convergence of $\{a_n\}$ is that, for every ϵ, a value of n can be found such that a_{n+m} lies between $a_n + \epsilon$, $a_n - \epsilon$, for $m = 1, 2, 3, \dots$. It follows from this that $\bar{a}_{n+m}$ lies between $\bar{a}_n + \bar{\epsilon}$, $\bar{a}_n - \bar{\epsilon}$; and this ensures the convergence of the sequence $\{\bar{a}_n\}$. Conversely, we see that if $\{\bar{a}_n\}$ is convergent, so also is $\{a_n\}$.

Next, let $\{\alpha_n\}$ be a convergent sequence of real numbers; then, between α_n and α_{n+1}, a real number $\bar{a}_n$ can be found which corresponds to a rational number a_n, determined as at the end of § 28. Conceive this to be done for every pair of consecutive elements in $\{\alpha_n\}$, and let us consider the sequences $\{\bar{a}_n\}$, $\{a_n\}$.

Since $\bar{a}_n - \bar{a}_{n+m} = (\bar{a}_n - a_n) + (a_n - a_{n+m}) + (a_{n+m} - \bar{a}_{n+m})$,
we have $|\bar{a}_n - \bar{a}_{n+m}| \leqq |\bar{a}_n - a_n| + |a_n - a_{n+m}| + |a_{n+m} - \bar{a}_{n+m}|$.
Now, corresponding to any real positive number δ, n may be so chosen that for every value of m, $|\bar{a}_n - a_n|$, $|a_n - a_{n+m}|$, $|a_{n+m} - \bar{a}_{n+m}|$ are each less than $\frac{1}{3}\delta$; hence, for such a value of n,

$$|\bar{a}_n - \bar{a}_{n+m}| < \delta, \text{ for } m = 1, 2, 3, \ldots,$$

thus the sequence $\{\bar{a}_n\}$ is convergent.

Again, $\{\alpha_n\} - \{\bar{a}_n\} = \{\alpha_n - \bar{a}_n\}$, and $|\alpha_n - \bar{a}_n| < |\alpha_n - \alpha_{n+1}|$, and, since $\{\alpha_n\}$ is convergent, n may be chosen so great that, for that and all higher values of n, all the differences $|\alpha_n - \alpha_{n+1}|$ are less than an arbitrarily fixed number, hence $|\alpha_n - \bar{a}_n|$ satisfies the same condition, and therefore the two convergent sequences $\{\alpha_n\}$, $\{\bar{a}_n\}$ satisfy the condition of equality; or they represent the same one of the new objects. It has been shewn above that, since $\{\bar{a}_n\}$ is convergent, so also is $\{a_n\}$. Now $\{a_n\}$ corresponds to a single real number a; therefore, to any convergent sequence $\{\alpha_n\}$, of which the elements are real numbers, there corresponds a real number a.

We have further to shew that, if $\{\alpha_n\}$, $\{\beta_n\}$ are two convergent sequences of real numbers, and a, b the corresponding real numbers, as just determined, then $a \gtreqless b$, according as $\{\alpha_n\} \gtreqless \{\beta_n\}$.

We know that $a \gtreqless b$, according as $\{a_n\} \gtreqless \{b_n\}$, where $\{b_n\}$ denotes the sequence of rational numbers which defines b, in the same way as $\{a_n\}$ defines a. Now $\{\alpha_n\} - \{\beta_n\} = \{\alpha_n - \beta_n\}$, and

$$\alpha_n - \beta_n = (\alpha_n - \bar{a}_n) + (\bar{b}_n - \beta_n) + (\bar{a}_n - \bar{b}_n);$$

and we can choose n so large that $|\alpha_n - \bar{a}_n|$ and $|\bar{b}_n - \beta_n|$ are each less than $\frac{1}{2}\eta$, where η is an arbitrarily chosen real positive number: therefore we see that $\alpha_n - \beta_n$ lies between $\bar{a}_n - \bar{b}_n + \eta$ and $\bar{a}_n - \bar{b}_n - \eta$. It follows easily that $\{\alpha_n\} \gtreqless \{\beta_n\}$, according as $\{\bar{a}_n\} \gtreqless \{\bar{b}_n\}$, or according as $\{a_n\} \gtreqless \{b_n\}$, and hence, as shewn above, according as $a \gtreqless b$. It has now been shewn that the objects which are represented by convergent sequences of real numbers have the same ordinal relation to one another as the real numbers to which those sequences have been shewn to correspond.

It appears, from what has now been proved, that, *to every convergent sequence of real numbers there corresponds a real number which may be taken to be defined by means of that sequence.*

There does not necessarily exist any rational number which corresponds in the same sense to a convergent sequence of rational numbers. The property of the aggregate of real numbers here stated embodies the characteristic difference between that aggregate and the aggregate of rational numbers; for the latter does not possess the corresponding property. It is this property of the aggregate of real numbers which makes it suitable to be the field of the real variable in the Theory of Functions.

THE ARITHMETICAL THEORY OF LIMITS

30. *If* $x_1, x_2, x_3, \dots x_n, \dots$ *is a sequence of real numbers such that a number* x *exists which has the property that, corresponding to any arbitrarily chosen positive number* ϵ, *a value of* n *can be found such that* $|x - x_n|$, $|x - x_{n+1}|$, $|x - x_{n+2}|$, ... *are all less than* ϵ, *then the number* x *is said to be the limit of the sequence* $x_1, x_2, \dots x_n, \dots$. *This fact may be denoted by the equation* $x = \lim\limits_{n \sim \infty} x_n$.

This definition is known as the arithmetical definition of a limit, and was first given*, in a form substantially identical with the above, by John Wallis.

It will be observed that the above definition contains no assertion as to the necessary existence of a limit of a sequence of numbers, but contains only a statement as to the relation of the limit to the numbers of the sequence, in case that limit exists.

There cannot be two numbers which both satisfy the condition of being a limit of the same sequence. For, if possible, let x, x' be two such numbers and let $|x - x'| = \delta$. Choose a value of ϵ, less than $\frac{1}{2}\delta$; then numbers n, n' can be found such that $|x - x_{n+m}|$, $|x' - x_{n'+m}|$, for all values 0, 1, 2, 3, ... of m, are less than ϵ. Suppose $n > n'$, then $|x - x_n|$ and $|x' - x_n|$ are both less than ϵ; hence $|x - x'| < 2\epsilon < \delta$, which is contrary to the condition $|x - x'| = \delta$.

It will now be shewn that, if the numbers of the sequence $\{x_n\}$ are real numbers, and if the sequence is a convergent one, then the real number x defined in the manner explained in § 29, by the sequence $\{x_n\}$, is the limit of the sequence.

For the two sequences $\{x_n\}$, $\{x\}$ both define the same number x, and therefore satisfy the condition of equality, which is that $|x - x_{n+m}| < \epsilon$, for any arbitrarily chosen ϵ, provided n be sufficiently great; and this is the condition that x should be the limit of the sequence $\{x_n\}$. A sequence of real numbers which has a limit must be convergent. For, if x is the limit of $\{x_n\}$, then for a sufficiently large value of n,

---

* *Arithmetica Infinitorum* (1655), Prop. 43, Lemma. See M. Cantor's *Geschichte der Mathematik*, Leipzig (1892), vol. II, p. 823.

$$| x - x_n |, \ | x - x_{n+1} |, \dots \ | x - x_{n+m} |, \dots$$

are all less than $\frac{1}{2}\epsilon$, where ϵ is arbitrarily chosen; now

$$| x_n - x_{n+m} | \leq | x - x_n | + | x - x_{n+m} |;$$

hence $| x_n - x_{n+m} | < \epsilon$, which is the condition of convergence of $\{x_n\}$.

As the complete result we have now the theorem known as the General Principle of Convergence*:

The necessary and sufficient condition that a sequence $x_1, x_2, \dots x_n, \dots$ *of real numbers may have a limit is that, corresponding to every arbitrarily chosen positive number* ϵ, *a value of* n *can be found such that* $x_n - x_{n+1}$, $x_n - x_{n+2}$, $x_n - x_{n+3}$, … *shall be all numerically less than* ϵ.

This theorem, which contains the criterion for the existence of a limit, as defined in accordance with the arithmetical definition of a limit, is a deduction from Cantor's theory of real numbers.

31. If the numbers of a sequence $\{x_n\}$ are rational numbers, instead of real numbers, the definition of the limit is applicable, and it is a necessary, but not a sufficient, condition for the existence of the limit, that the sequence should be convergent. Strictly speaking, if a convergent sequence of rational numbers has a limit, that limit is also a rational number; but from the existence of convergent sequences of rational numbers which have no limit there arises the necessity for the extension of the domain of number, so that in the extended domain every convergent sequence may have a limit; this extension has been carried out by substituting Real Number for Rational Number. However, although a convergent sequence of rational numbers which has no rational limit, has in this strict sense no limit at all, by reason of the convergent sequence of those real numbers which correspond to the rational numbers having an irrational number as limit, and since, as has been seen above, these real numbers are for practical purposes not distinguished from the rational numbers to which they correspond, it is usual to consider this irrational number to be the limit of the sequence of rational numbers. We may thus assert that any convergent sequence of rational numbers which has not a rational number as limit has an irrational number as its limit. This assertion is a correct one for the practical purposes of Mathematical Analysis.

32. The method of limits, which is essential both to pure Analysis and to the applications of Analysis in Geometry and in Kinetics, had a geometrical origin in the Method of Exhaustions, which was applied by the Greek geometers to determine lengths, areas, and volumes, in simple cases.

---

* This term "das allgemeine Convergenzprinzip" is due to P. Du Bois-Reymond; see his *Allgemeine Functionentheorie*, Tübingen (1882).

This method, supplemented by the notion of the numerically *infinite*, was developed in later times, in various forms, into a general method which formed the basis of the Infinitesimal Calculus. The traditional geometrical conception of a limit may be exemplified by the case of the determination of the length of a curve as the limit of a sequence of properly chosen inscribed polygons. The lengths of the perimeters of the polygons are regarded as continually approaching the required length of the curve, whilst the number of sides of the polygons is continually increased, and the maximum length of the sides of a polygon is diminished indefinitely. The limit, the length of the curve, is then regarded as actually reached at the end of a process described as making the number of sides of the polygon infinite; this mode of attainment of the limit being however inaccessible to the sensuous imagination, and disguising an actual qualitative change of a geometrical figure, which possesses corners and is bounded by segments of straight lines, into one which has no corners and has a curvilinear boundary. No doubt was felt as to the existence of the limit, which was regarded as obvious from geometrical intuition. That a curve possesses a length, or an area, was considered to require no proof. The first mathematician who recognized the necessity for a proof of the existence of a limit was Cauchy, who gave a proof of the existence of the integral of a continuous function. That the logical basis of the traditional method of limits is defective has in recent times received *a posteriori* confirmation by the exhibition of continuous functions which possess no differential coefficient, and by many other cases of exception to what were regarded as ordinary results of analysis resting on the method of limits, which have been brought to light by those mathematicians who have been engaged in examining the foundations of analysis.

The arithmetical theory of limits, which is summed up in the general principle of Convergence, provides a definite criterion for the existence of the limit of a sequence of numbers; and a considerable part of modern analysis is concerned with obtaining special forms of the general criterion adapted for use in special classes of cases. The theory is essentially dependent upon the theory of irrational numbers; for, in default of an arithmetical theory of irrational numbers, all attempts to prove* the existence of a limit of a convergent sequence are doomed to inevitable failure; and this for the simple reason that a convergent sequence of rational numbers does not necessarily possess a limit which is within the domain of such numbers. The definition of real numbers by means of convergent sequences of rational numbers is not a mere postulation of the existence of limits to such sequences; it involves rather the introduction of an enlarged conception of number, of such a character that the scheme of ordered real numbers should form a consistent whole, and such that every convergent sequence of numbers in the domain of real number *necessarily* has a limit within that domain. The postulation of the existence of the aggregate of real numbers is justified by shewing that a complete scheme of definitions and postulates can be set up for the elements of this aggregate, and that such a scheme does not lead to contradiction*. As regards the existence of limits in the case of lengths, areas, volumes, &c., referred to above, the order of procedure is a reversal of the traditional one, the existence of the limit being no longer inferred from geometrical intuition. For example, in the case of the determination of the length of a curve, that length is not assumed to be independently known to exist, but is defined as the arithmetical limit of the sequence of numbers which represent the perimeters of a suitable sequence of inscribed polygons. When this sequence is convergent, and its limit is independent of the particular choice of the polygons, subject to a suitable restriction, then the limit so obtained determines the length of the curve. In case no such limit exists, the curve is regarded as not having a length.

---

* An interesting discussion of various methods which have been suggested of proving the existence of a limit will be found in Du Bois-Reymond's *Allgemeine Functionentheorie*

EQUIVALENCE OF THE DEFINITIONS OF DEDEKIND AND CANTOR

33. In order to establish the equivalence of the definitions of irrational numbers, as given by Dedekind and by Cantor, it must be shewn that every convergent sequence of rational numbers defines uniquely a section of all the rational numbers, and that this section is the same for all convergent sequences which represent the same real number in accordance with rule (1) in § 26. Conversely, it must be shewn that any number defined by a section can also be represented by a convergent sequence of rational numbers.

To shew that, corresponding to the convergent sequence $\{x_n\}$ which, in accordance with the Cantor theory, defines the real number x, a section can be found:—Let r be any rational number, and let $\bar{r}$ be the corresponding real number represented by $\{r\}$. The real number $x - \bar{r}$ is represented by $\{x_n - r\}$; and if this number is not zero, then (see § 28), from and after some fixed value of n, $x_n - r$ has a fixed sign, positive or negative according to the value of r. A section of the rational numbers may now be defined as follows:—Let every number r such that $x_n - r$ is negative, from and after some fixed value of n, be placed in the class R_2; and let every number for which $x_n - r$ is positive, from and after some fixed value of n, be placed in the class R_1. If there exists a rational number r, such that neither of these cases arises, then $x \equiv \bar{r}$, and r may be put into either of the classes

---

* On this mode of regarding the aggregate of real numbers as dependent upon a complete consistent scheme of definitions and axioms, see Hilbert, "Ueber den Zahlbegriff," *Jahresber. d. deutsch. math. Vereinigung*, vol. VIII (1900), p. 180

R_1, R_2. It has thus been shewn that a section of the rational numbers can be determined, corresponding to the convergent sequence $\{x_n\}$.

Next, let $\{x_n'\}$ be any other convergent sequence which represents the same real number x, as $\{x_n\}$ does. We have to shew that the section of the rational numbers which corresponds to $\{x_n'\}$ is identical with that which corresponds to $\{x_n\}$. If, as before, r denote any rational number, we have $\{x_n - r\} \equiv \{x_n' - r\}$. Now a value of n can be found, from and after which $x_n - r$ and $x_n' - r$ both have fixed signs independent of n, and they must have the same sign. It follows that a number r which belongs to the class R_1, must also belong to the class R_1', by which the section corresponding to $\{x_n'\}$ is defined; and also a number r which belongs to the class R_2, necessarily belongs to R_2', except in the case $\{x_n\} = \{x_n'\} = \bar{r}$. It has thus been shewn that the section (R_1, R_2) which corresponds to $\{x_n\}$ is identical with the section (R_1', R_2') which corresponds to $\{x_n'\}$.

34. To shew that a convergent sequence can always be found such as to define the number corresponding to a given section (R_1, R_2), we observe that two rational numbers can always be found, one of which is in R_1, and the other in R_2, and such that their difference is numerically less than a given arbitrarily small rational number ϵ. Let A be any rational number in R_1, and let ϵ' be a rational number $< \epsilon$. Then of the numbers $A + \epsilon'$, $A + 2\epsilon'$, ... $A + r\epsilon'$, ... there must be a last one $A + r\epsilon'$ which falls in R_1, for $A + n\epsilon'$ may be made as large as we please by taking n large enough: the next number $A + (r + 1)\epsilon'$ is then in R_2; and these numbers $A + r\epsilon'$, $A + (r + 1)\epsilon'$, whose difference is $\epsilon' < \epsilon$, are the two numbers required. Moreover, if B is a rational number in R_2, the two numbers may be so determined that both lie between A and B; for we need only take ϵ' to be of the form $\frac{1}{s}(B - A)$, where s is a positive integer so chosen that $\frac{1}{s}(B - A) < \epsilon$.

Now let $\{\epsilon_n\}$ be any convergent aggregate of rational numbers, which has zero for its limit. Determine x_1 in R_1, and x_2 in R_2, so that $x_2 - x_1 < \epsilon_1$; next take x_3 in R_1, and x_4 in R_2, so that $| x_3 - x_4 | < \epsilon_2$; and that x_3, x_4 both lie between x_1 and x_2. Proceeding in this way, we can determine x_{2n-1}, x_{2n} rational numbers of different classes, so that $| x_{2n-1} - x_{2n} | < \epsilon_n$; then either of the sequences $\{x_1, x_3, x_5, ...\}$, $\{x_2, x_4, ...\}$ defines the number which is represented by the section (R_1, R_2).

To prove this, we observe that $\{x_{2n-1}\}$ is a convergent sequence, since all the elements are $< x_2$, and $x_1 < x_3 < x_5 \dots$.

Again, suppose a is a rational number belonging to R_2, we can shew that, provided a rational number b exists in R_2 which is less than a, then a is greater than all the numbers $x_1, x_3, \dots$ by more than $a - b$. For

$$a - x_{2n-1} = (a - b) + (b - x_{2n-1}) > a - b,$$

however small $b - x_{2n-1}$ may become. Hence, unless a is the smallest rational number in R_2, the real number $\{a\}$ which corresponds to a is greater than the number $(x_1, x_3, \ldots)$.

Again, the sequences $\{x_{2n-1}\}$, $\{x_{2n}\}$ represent the same number, since their difference is the aggregate $\{\epsilon_n\}$ which defines zero. It now appears, by reasoning similar to the above, that any number a in R_1 is such that the real number $\{a\}$ is less than the number $\{x_{2n}\}$, unless a is the greatest rational number in R_1.

If either R_1 has a greatest rational number, or R_2 has a least one, the real number $\{a\}$ which corresponds to this rational number a, is itself defined by (R_1, R_2), and is the number represented by either of the sequences $\{x_{2n-1}\}$, $\{x_{2n}\}$. In any case, either of these two sequences defines the number given by the section (R_1, R_2).

The complete equivalence of the two theories of Dedekind and of Cantor has now been established. The first theory operates with the whole aggregate of rational numbers, the second with sequences selected out of that aggregate.

THE NON-EXISTENCE OF INFINITESIMALS

35. It should be remarked that, in assuming that every section of the aggregate of real numbers defines a single real number, it has been implicitly assumed that *if a, b are any two positive real numbers, such that $a < b$, then a positive integer n can be found such that $na > b$.*

This is the arithmetical analogue of the so-called principle of Archimedes.

If any real numbers existed which are ordinally greater than all the numbers $a, 2a, 3a, \ldots$, then a section of the aggregate of real numbers would be defined by considering all numbers greater than all the numbers $a, 2a, 3a, \ldots$ to be in one class, and all the remaining real numbers to be in the other class; and this section would define a real number N. If now ϵ be an arbitrarily chosen positive number less than a, then $N - \epsilon$ is a number which is less than some of the numbers $a, 2a, 3a, \ldots$; and there must be a first of this set of numbers such that $N - \epsilon$ is less than it. Let this be pa; thus $N - \epsilon < pa$, hence $N < pa + \epsilon < (p + 1)a$; which is contrary to the hypothesis that no number na is in the class of numbers which are $> N$.

The property of the aggregate of real numbers which has been established may be denoted by the statement that *the aggregate of real numbers forms an Archimedean system*; and this property of the aggregate is essentially equivalent to the property that every section of the aggregate defines a single number of the aggregate.

A consequence of the fact that the aggregate of real numbers forms an Archimedean system is that so-called infinitesimal numbers do not exist within the aggregate. Every positive number ϵ, being such that an integer n can be found such that $n\epsilon > 1$, is a finite number, in the sense in which finite numbers were distinguished from infinitesimals in the older forms of the Infinitesimal Calculus. *In Arithmetical Analysis the conception of the actually infinitesimal has no place.* When the expression "infinitesimal" is used at all, it is to describe the process by which a variable to which the numbers of a sequence converging to zero are successively ascribed, as values, approaches the limit zero; thus an infinitesimal is a variable in a state of flux, never a number. Such a form of expression, appealing as it does to a mode of thinking which is essentially non-arithmetical, is better avoided.

THE THEORY OF INDICES

36. When m is a positive integer, and x a rational number, x^m was defined to denote $x \times x \times x \times \ldots \times x$ (m factors); and this definition may be extended to the case in which x is any number defined by a convergent sequence; so that if x is defined by $\{x_n\}$, x^m is defined by $\{x_n{}^m\}$. It thus appears that for any real number x, we have, provided m and n are positive integers, $x^m \times x^n = x^{m+n}$.

If we assume x^0 and x^{-m} to be defined as having such a meaning that this law of indices holds when m or n is zero, or a negative integer, we can at once interpret x^0 and x^{-m}; for

$$x^0 \times x^n = x^{n+0} = x^n, \text{ thus } x^0 = 1;$$

and
$$x^{-n} \times x^n = x^0 = 1, \text{ thus } x^{-n} = \frac{1}{x^n}.$$

When p/q is a rational fraction, we shall define $x^{\frac{p}{q}}$ to have such a meaning that the above law of indices holds when either or both of m, n may be rational fractions. With this assumption

$$x^{\frac{p}{q}} \times x^{\frac{p}{q}} \times x^{\frac{p}{q}} \times \ldots \times x^{\frac{p}{q}} \ (q \text{ factors}) = x^p;$$

hence $(x^{\frac{p}{q}})^q = x^p$; or $x^{\frac{p}{q}}$ is, if it exists, a number whose qth power is x^p. The problem of determining, if possible, a number $x^{\frac{p}{q}}$, is that of finding a number whose qth power is a given number; and it has been already shewn that this is not always a possible operation within the domain of rational numbers.

It will now be shewn that, in the domain of real numbers, *the operation of finding $x^{\frac{p}{q}}$ is always a possible one when x is positive; and also when x is* *negative, provided however that, in this latter case, q is an odd number, or if it is even, p is not odd.*

The following lemma will be required:—*If α is any real positive number less than unity, a positive integer m can be found such that $\alpha^m < \epsilon$, where ϵ is an arbitrarily prescribed positive number, or in other words,* $\lim\limits_{n \to \infty} \alpha^n = 0$.

Since $\alpha^n > \alpha^{n+1}$, the sequence $(\alpha, \alpha^2, \dots \alpha^n, \dots)$ is convergent.

Suppose, if possible, that the sequence represents a positive number k different from zero; then m may be so chosen that $\alpha^m, \alpha^{m+1}, \dots$ all differ from k by less than the arbitrarily prescribed number δ, say $\alpha^m = k + \eta$, where $\eta < \delta$. We have therefore $\alpha^{m+1} = (k + \eta)\alpha < (k + \delta)\alpha$; now δ can be chosen to be equal to $\frac{k(1-\alpha)}{1+\alpha}$, then $\alpha^{m+1} < k - \delta$; and this is contrary to the condition imposed in the choice of m.

It follows that k cannot be different from zero; and thus the lemma is established.

Suppose now that a is any positive number, rational or not, which lies between N^q and $(N+1)^q$, where N is a positive integer; we shall first shew that a number $N + h$, where $h < 1$, can always be found such that $a - (N+h)^q$ is positive, and less than $a - N^q$. We find by division $(N+h)^q - N^q = \{(N+h) - N\}\{(N+h)^{q-1} + (N+h)^{q-2}N + \dots + N^{q-1}\}$; hence, if h is positive and less than unity, $(N+h)^q - N^q$ lies between

$$qhN^{q-1} \text{ and } qh(N+1)^{q-1}.$$

Since $$a - (N+h)^q = (a - N^q) - \{(N+h)^q - N^q\},$$

we must take h not greater than $\frac{a - N^q}{q(N+1)^{q-1}}$, in order that $a - (N+h)^q$ may certainly be positive; and the difference $a - (N+h)^q$ is then less than

$$(a - N^q) - qhN^{q-1}.$$

Let $$h = \frac{a - N^q}{q(N+1)^{q-1}},$$

then $$a - (N+h)^q < (a - N^q)\left\{1 - \left(\frac{N}{N+1}\right)^{q-1}\right\}.$$

Let $N_1 = N + h$, then N_1 is such that

$$a - N_1^q < (a - N^q)\left\{1 - \left(\frac{N}{N+1}\right)^{q-1}\right\},$$

and $$N_1 > N.$$

In a similar manner, we can shew that a number N_2 exists which is $> N_1$, and such that

$$a - N_2^q < (a - N_1^q)\left\{1 - \left(\frac{N_1}{N_1+1}\right)^{q-1}\right\}.$$

Proceeding in this manner, we obtain a series of numbers $N, N_1, N_2, \ldots N_r, \ldots$ such that $N_r > N_{r-1}$, and that $a - N_r^q$ is positive, and less than

$$(a - N_{r-1}^q)\left[1 - \left(\frac{N_{r-1}}{N_{r-1}+1}\right)^{q-1}\right].$$

We shall now shew that $(N, N_1, N_2, \ldots N_r, \ldots)$ is a convergent sequence which defines a number whose qth power is a.

The sequence $\{N_r\}$ is convergent, since $N_r > N_{r-1}$, and every N_r is less than $N + 1$. The qth power of the number defined by this convergent sequence is $\{N_r^q\}$, and we shall shew that this defines the number a or $\{a\}$.

We have

$$a - N_r^q < (a - N^q)\left[1 - \left(\frac{N_{r-1}}{N_{r-1}+1}\right)^{q-1}\right]\left[1 - \left(\frac{N_{r-2}}{N_{r-2}+1}\right)^{q-1}\right]\cdots$$
$$\left[1 - \left(\frac{N}{N+1}\right)^{q-1}\right] < (a - N^q)\left[1 - \left(\frac{N}{N+1}\right)^{q-1}\right]^r$$

for
$$\frac{N}{N+1} < \frac{N_{r-s}}{N_{r-s}+1},$$

and hence
$$1 - \left(\frac{N}{N+1}\right)^{q-1} > 1 - \left(\frac{N_{r-s}}{N_{r-s}+1}\right)^{q-1}.$$

Now $1 - \left(\frac{N}{N+1}\right)^{q-1}$ is a proper fraction, hence from the lemma proved above we infer that a power r of the expression can be found which is less than an arbitrarily chosen positive number; which number we may take to be $\frac{\epsilon}{a - N^q}$. Hence, corresponding to every ϵ, a number r can be found such that $a - N^q_{r+s} < \epsilon$, for $s = 0, 1, 2, \ldots$, and therefore the sequence $\{N_r^q\}$ defines the number $\{a\}$ or a.

If a is a positive proper fraction, we have $(a^2)^q < a$; hence we may take N to be equal to a^2, instead of to a positive integer Then $a < (N + 1)^q$; thus this value of N will play the same part as the integral value in the above proof, and the reasoning is the same as before.

37. It has now been shewn that in every case a real number can be found of which the qth power is a given positive number a. It thus appears that $x^{\frac{p}{q}}$ has an interpretation within the domain of real numbers, when x is any positive number, and $\frac{p}{q}$ is a positive rational fraction.

We interpret $x^{-\frac{p}{q}}$ to be such that

$$x^{-\frac{p}{q}} \times x^{\frac{p}{q}} = x^0 = 1,$$

or
$$x^{-\frac{p}{q}} = 1/x^{\frac{p}{q}}.$$

If x is a negative number $-x'$, we have $(-x')^{\frac{p}{q}}$, defined as a number whose qth power is $(-x')^p$, and $(-x')^p$ is x'^p or $-x'^p$, according as p is even or odd.

If p is even, $(-x')^{\frac{p}{q}}$ can be interpreted as the value of $x'^{\frac{p}{q}}$. If p is odd, and q is odd, $(-x')^{\frac{p}{q}}$ may be interpreted as $-x'^{\frac{p}{q}}$. When p is odd, and q is even, we have obtained no interpretation of $(-x')^{\frac{p}{q}}$.

To complete the theory of indices in such a way that $(-x')^{\frac{2r+1}{2s}}$ may have an interpretation, we should require a further extension of the conception of number. This further extension takes place by the introduction of complex number, which is however outside the limits imposed upon this work as a treatise dealing only with real number.

38. The only case in which x^n, for a positive x, has not been defined, is that in which n is not a rational number. To extend the definition to this case, we suppose n to be defined by a convergent sequence $\{n_r\}$, in which all the numbers n_r are rational. We shall shew that the aggregate $\{x^{n_r}\}$ is convergent, and the number which it defines we shall denote by x^n.

We have $x^{n_r} - x^{n_{r+s}} = x^{n_r}\{1 - x^{n_{r+s}-n_r}\}$; now, since $\{n_r\}$ is a convergent aggregate, all the numbers n_r are numerically less than some fixed number, and therefore $| x^{n_r} | < A$, where A is some fixed number.

First suppose $x > 1$, then

$$x^{n_r} - x^{n_{r+s}} = x^{n_{r+s}}(x^{n_r - n_{r+s}} - 1) = x^{n_r}(1 - x^{n_{r+s}-n_r}),$$

hence $$| x^{n_r} - x^{n_{r+s}} | < A \, | x^{|n_r - n_{r+s}|} - 1 | .$$

Now let r be so chosen that, for all values of s,

$$| n_r - n_{r+s} | < \frac{1}{q},$$

where q is a positive integer; then

$$x^{|n_r - n_{r+s}|} - 1 < x^{\frac{1}{q}} - 1 < \frac{x-1}{1 + x^{\frac{1}{q}} + x^{\frac{2}{q}} + \dots + x^{\frac{q-1}{q}}},$$

hence $$| x^{|n_r - n_{r+s}|} - 1 | < \frac{x-1}{q},$$

or $$| x^{n_r} - x^{n_{r+s}} | < A \frac{x-1}{q},$$

and if q is chosen so that $\frac{1}{q} < \frac{\epsilon}{A(x-1)}$, where ϵ is a fixed number, we see that r may be so chosen that $| x^{n_r} - x^{n_{r+s}} | < \epsilon$, for all values of s; therefore $\{x^{n_r}\}$ is a convergent sequence.

If $x < 1$, then $\left\{\frac{1}{x^{n_r}}\right\}$ is a convergent sequence, and therefore $\{x^{n_r}\}$ is also convergent, since it is the quotient of $\{1\}$ and $\{x^{n_r}\}$. If $x = 1$, then $\{x^{n_r}\} = 1$. Thus in every case $\{x^{n_r}\}$ is a convergent sequence if $\{n_r\}$ is convergent.

It is easily seen that, if $\{n_r'\}$ is any other convergent sequence which also converges to n, the two sequences $\{x^{n_r}\}$, $\{x^{n_r'}\}$ both define the same number. Thus the value of x^n is independent of the particular sequence of rational numbers employed in defining n.

Since $\{x^{n_r}\} \times \{x^{m_r}\} = \{x^{n_r+m_r}\}$, we see that the definition of x^n, when n is not rational, is such that the relation $x^m \times x^n = x^{m+n}$ is satisfied.

THE REPRESENTATION OF REAL NUMBERS

39. The ordinary mode of representation of a real number is by means of a decimal, or more generally by a radix-fraction. When the decimal is non-terminating, this mode of representation is a case of the representation by a convergent sequence of rational numbers, in accordance with Cantor's theory. For example, the number π is represented by the sequence

$$(3,\ 3{\cdot}1,\ 3{\cdot}141,\ 3{\cdot}1415,\ 3{\cdot}14159,\ 3{\cdot}141592,\ \ldots),$$

where, by known processes, any prescribed element can be found as the result of a definite number of arithmetical operations.

The general theorem will be established that *every positive non-rational real number N is uniquely representable by means of a non-terminating series of radix-fractions, of which r, the radix, is any integer* ≥ 2.

Of the numbers $0, r, 2r, 3r, \ldots$, there is (see § 35), of all those which are less than rN, a greatest one $c_0 r$, which may be zero; thus

$$rN > c_0 r, \text{ and } < (c_0 + 1)\, r;$$

it follows that
$$N = c_0 + \frac{N_1}{r},$$

where N_1 is a positive number less than r.

In a similar manner we obtain

$$N_1 = c_1 + \frac{N_2}{r},\quad N_2 = c_2 + \frac{N_3}{r},\ \ldots\ N_n = c_n + \frac{N_{n+1}}{r},$$

where $N_2, N_3, \ldots N_{n+1}$ are all $< r$; therefore

$$N = c_0 + \frac{c_1}{r} + \frac{c_2}{r^2} + \ldots + \frac{c_n}{r^n} + \frac{N_{n+1}}{r^{n+1}},$$

where $c_0, c_1, c_2, \ldots c_n$ are each of them positive integral, or zero, and $0 < N_{n+1} < r$.

Since
$$N - \left(c_0 + \frac{c_1}{r} + \frac{c_2}{r^2} + \ldots + \frac{c_n}{r^n}\right) < \frac{1}{r^n},$$

and it has been shewn that $\frac{1}{r^n}$ has the limit zero, as n is indefinitely increased, we see that the sequence, of which the nth element is

$$c_0 + \frac{c_1}{r} + \frac{c_2}{r^2} + \dots + \frac{c_n}{r^n},$$

is convergent, and represents the real number N. This is expressed by

$$N = c_0 + \frac{c_1}{r} + \frac{c_2}{r^2} + \dots + \frac{c_n}{r^n} + \dots,$$

in which N is represented by a non-terminating radix-fraction.

Let us now consider the case in which N is a rational number $\frac{a}{b}$, in its lowest term. We have $a = \alpha_0 b + \beta_0$, where $\beta_0 < b$; and $r\beta_0 = \alpha_1 b + \beta_1$, where $\beta_1 < b$; $r\beta_1 = \alpha_2 b + \beta_2, \dots, r\beta_{n-1} = \alpha_n b + \beta_n$, where $\beta_1, \beta_2, \dots \beta_n$ are all less than b.

If one of the numbers β, say β_n, is zero, we have

$$N \equiv \frac{a}{b} = \alpha_0 + \frac{\alpha_1}{r} + \frac{\alpha_2}{r^2} + \dots + \frac{\alpha_n}{r^n};$$

and thus N is expressed in terminating radix-fractions; this case can only arise when b contains only prime factors of r. The terminating series of radix-fractions can be replaced by a periodic one which does not terminate. For if we use $\alpha_n - 1$ instead of α_n, as the numerator of r^n, we have

$$r\beta_{n-1} = (\alpha_n - 1)\, b + b;$$

thus β_n becomes b instead of zero, and

$$rb = (r-1)\, b + b;$$

thus $\beta_n, \beta_{n+1}, \dots$ are all equal to b; and $\alpha_{n+1}, \alpha_{n+2}, \dots$ are all equal to $r - 1$. Thus N is represented by

$$N \equiv \frac{a}{b} = \alpha_0 + \frac{\alpha_1}{r} + \frac{\alpha_2}{r^2} + \dots + \frac{\alpha_n - 1}{r^n} + \frac{r-1}{r^{n+1}} + \frac{r-1}{r^{n+2}} + \dots.$$

It thus appears that *a rational number, which in its lowest terms has a denominator which contains only prime factors of r, is capable of a double representation;* (1) *by a terminating series of radix-fractions;* (2) *by a non-terminating series of radix-fractions, of which the numerators after some fixed one are all $r - 1$.*

In case none of the numbers $\beta_1, \beta_2, \dots \beta_n, \dots$ vanishes, it is clear that since all these numbers are either $1, 2, 3, \dots b - 1$, they cannot be all unequal. Suppose β_n is the first which is repeated, and let $\beta_n = \beta_{n+m}$; it is then clear that $\beta_{n+1} = \beta_{n+m+1}, \beta_{n+2} = \beta_{n+m+2}, \dots$; and therefore the number is represented by a recurring series of radix-fractions.

40. When a number is defined by means of a convergent sequence of some special form, it is in general not immediately obvious whether the number is rational or irrational. Many special investigations relating to particular cases, and various general criteria, have been given by well-known mathematicians.

One of the most important modes of such representation of a number is that by an endless continued fraction. This fraction may be regarded as an aggregate, each element of which is a finite continued fraction. Legendre established the fundamental theorem that a number represented by an endless continued fraction

$$\frac{a_1}{b_1 \pm}\,\frac{a_2}{b_2 \pm}\,\frac{a_3}{b_3 \pm}\cdots\frac{a_n}{b_n \pm}\cdots,$$

that is, by an aggregate of which the nth element is

$$\frac{a_1}{b_1 \pm}\,\frac{a_2}{b_2 \pm}\,\frac{a_3}{b_3 \pm}\cdots\frac{a_n}{b_n},$$

is irrational*, provided the positive integers a_n, b_n are such that for every value of n, $b_n - a_n \geqq 1$; except that when $b_n - a_n = 1$, for every value of $n \geq m$, where m is some fixed number, and when at the same time the signs before all the fractions $\frac{a_n}{b_n}$, for $n > m$, are negative, then the continued fraction converges to unity, or to a rational fraction, according as $m = 1$, or $m > 1$.

This theorem contains as special cases the theorems previously established by Lambert, that e^x, $\tan x$, $\log_e x$, $\tan^{-1} x$, π are irrational for rational values of x. The irrationality of e and e^2 was first proved by Euler†. Legendre‡ himself applied the general theorem to prove the irrationality of π^2, although his proof was lacking in rigour.

The following general theorem has been proved§ by Cantor.

If b, b', b'', ... form a set of positive integers such that, q being any arbitrarily chosen integer, all the numbers $1, b, bb', bb'b'', \ldots$, from and after some fixed number of the sequence, are divisible by q; then any number N can be uniquely represented by

$$I + \frac{\lambda}{b} + \frac{\mu}{bb'} + \frac{\nu}{bb'b''} + \ldots,$$

where I is an integer, and λ, μ, ν, ... are integers (including 0) such that

$$\lambda \leq b - 1, \quad \mu \leqq b' - 1, \quad \nu \leqq b'' - 1, \ldots.$$

Further, in order that the number N may be rational, it is necessary that, from and after some fixed term of the series, all the numbers $\lambda, \mu, \nu, \ldots$ have their highest possible values. If this condition is not satisfied, N is irrational.

---

* A proof of this theorem is given by Pringsheim, "Ueber die Convergenz unendlicher Kettenbrüche," *Sitzungsberichte d. bayer. Akad.* vol. xxvii (1897), p. 318.

† On the history of these theorems see Pringsheim's article "Ueber die ersten Beweise der Irrationalitat von e und π," *Sitzungsberichte d. bayer. Akad.* vol. xxvii (1897).

‡ See his *Éléments de Géométrie*, Note 4; see also Rudio's work, "Archimedes, Huygens, Lambert, Legendre" (1892), p. 166.

§ Schlömilch's *Zeitschrift*, vol. xiv (1869), p. 121, "Ueber die einfachen Zahlensysteme."

As an example of this theorem, the number e represented by

$$2+\frac{1}{2}+\frac{1}{2.3}+\frac{1}{2.3.4}+\ldots$$

is seen to be irrational.

Another mode of representation is that in which the sequence of integers $b, b', b'', \ldots$, from a particular element onwards, is periodic. In this case, the necessary and sufficient condition that the number represented by

$$\frac{\beta}{b}+\frac{\beta'}{bb'}+\frac{\beta''}{bb'b''}+\ldots$$

should be rational, is that the sequence $\beta, \beta', \beta'', \ldots$ be, from and after some fixed number of the sequence, periodic. This is a generalization of the theorem relating to a number represented by radix-fractions.

If $b=2, b'=3, b''=4, \ldots$, we obtain the theorem* that the number represented by

$$\frac{c_1}{2!}+\frac{c_2}{3!}+\frac{c_3}{4!}+\ldots+\frac{c_n}{n!}+\ldots,$$

where $c_n \leq n-1$, is rational, only if, from and after some particular value of n, $c_n = n-1$.

A mode of representation of numbers by sequence of products has been given† by Cantor. He shews that every number $N > 1$, can be uniquely represented in the form

$$\left(1+\frac{1}{a}\right)\left(1+\frac{1}{b}\right)\left(1+\frac{1}{c}\right)\left(1+\frac{1}{d}\right)\ldots,$$

where $a, b, c, \ldots$ are integers such that

$$b \geqq a^2, \quad c \geqq b^2, \quad d \geqq c^2, \ldots.$$

The number a is determined as the integral part of $\frac{N}{N-1}$. If $\frac{Na}{a+1}=B$, b is the integral part of $\frac{B}{B-1}$; if $\frac{Bb}{b+1}=C$, c is the integral part of $\frac{C}{C-1}$; and so on.

As an example, $\sqrt{2}$ is represented by

$$\left(1+\frac{1}{3}\right)\left(1+\frac{1}{17}\right)\left(1+\frac{1}{577}\right)\left(1+\frac{1}{665857}\right)\ldots,$$

where $17 = 2.3^2-1, \quad 577 = 2.17^2-1, \quad 665857 = 2.577^2-1, \ldots$.

---

* See Stéphanos, *Bulletin de la soc. math. de France*, vol. VII (1879). For further information on the history of this subject see Pringsheim's article I. A. 3, in the *Encyclopädie der math. Wissenschaften*.

† Schlömilch's *Zeitschrift*, vol. XIV (1869), p. 152, "Zwei Satze...."

The criterion for determining whether N is rational or irrational is the following:

The number represented by

$$\left(1+\frac{1}{a}\right)\left(1+\frac{1}{b}\right)\left(1+\frac{1}{c}\right)\ldots,$$

where $$b \geq a^2, \quad c \geq b^2, \ldots.$$

all the numbers $a, b, c, \ldots$ being positive integers, is rational if, from and after some fixed number of the sequence $a, b, c, \ldots$, each number is the square of the preceding number of the sequence; but the number is irrational if this condition is not satisfied.

THE CONTINUUM OF REAL NUMBERS

41. If a_1, b_1 are any two real numbers such that $a_1 < b_1$, then two real numbers a_2, b_2, $(a_2 < b_2)$, can be found both lying between a_1, b_1, and such that the difference between a_2, b_2 is as small as we please, *i e* $b_2 - a_2 < \epsilon$, where ϵ is an arbitrarily prescribed number. Between a_2, b_2, two more numbers a_3, b_3, $(a_3 < b_3)$, can be found whose difference is again as small as we please; and this process may be carried on indefinitely. This property of the aggregate of real numbers may be expressed, to use the term introduced by G. Cantor, by saying that the aggregate of real numbers is *connex*; it arises from the fact that an indefinite series of numbers can be found which lie between any two given numbers. If we anticipate a term which will be introduced when we come to the general theory of aggregates, the property of connexity may be expressed by saying that *the aggregate of real numbers is everywhere dense.*

It will further be observed that the aggregate of rational numbers is also connex; so that, as far as this property is concerned, there is nothing to differentiate the one aggregate from the other.

If the difference of a_n and b_n is denoted by ϵ_n, and the sequence $\epsilon_1, \epsilon_2, \ldots \epsilon_n, \ldots$ satisfies the condition that, corresponding to any fixed arbitrarily small positive number η, a value of n can be found such that $\epsilon_n, \epsilon_{n+1}, \ldots$ are all less than η, then there exists a single real number x which is greater than all the numbers $a_1, a_2, \ldots$, and less than all the numbers $b_1, b_2, \ldots$. This number x is the limit of either of the sequences $(a_1, a_2, \ldots a_n, \ldots)$ and $(b_1, b_2, \ldots b_n, \ldots)$, and is defined by a section of all the real numbers.

If we confine ourselves to the domain of rational numbers, there subsists in that domain no such property; that is, the above numbers a, b being all rational, no such rational number as x necessarily exists.

In the domain of Real Number, (a), every convergent sequence has a limit which is a number belonging to the domain, and (b), every number is the limit of properly chosen sequences of numbers belonging to the domain. The possession by the domain of real numbers of these properties (*a*) and (*b*) is expressed by saying that *the aggregate of real numbers is perfect.*

The domain of Rational Number possesses the property (*b*) but not the property (*a*); consequently the aggregate of rational numbers is not perfect.

From the point of view of Dedekind's theory, the property that the aggregate of real numbers is perfect expresses the fact that every section of the real numbers corresponds to a single real number, and the converse. A section of the rational numbers does not always correspond to a rational number; consequently the aggregate of rational numbers is not perfect

We here give the name *continuum** to an aggregate which possesses the two properties of being connex, and of being perfect. This is in the first instance taken to be the definition of the meaning of the word continuum, as it is frequently used in Analysis. Thus the aggregate of real numbers forms a continuum; whereas the aggregate of rational numbers is essentially discrete, and does not form a continuum, since one of the two essential properties of a continuum is absent.

The aggregate of real numbers is spoken of as the *continuum of real numbers*, or *the arithmetic continuum.*

The real numbers which lie between two numbers a, b do not form a continuum in accordance with the above definition; but if the two numbers a, b themselves are considered to be included in the total aggregate, then this completed aggregate does form a continuum.

It should be remarked that, in accordance with a somewhat different definition of the term continuum, employed by Weierstrass, and which will be referred to in Chapter II, the real numbers between a and b form a continuum.

All the real numbers x such that $a \leqq x \leqq b$, in the ordinal sense of the symbols $<, =, >$, are said to form an interval (a, b); and such an interval is frequently described as a closed interval.

The real numbers x which are such that $a < x < b$, are frequently said to form an *open* interval (a, b).

The closed interval (a, b) is a continuum, since it satisfies the two necessary conditions for the applicability of the term; but the open interval (a, b) is not a continuum in this sense of the term, as it contains convergent sequences which have no limit belonging to the open interval. Such an open interval has been termed by Cantor a *semi-continuum*, but, in accordance with the observation made above, it may be termed a Weierstrassian continuum.

---

* See Cantor, *Math. Annalen*, vol. XXI (1883), p. 576.

Of the two essential properties of the arithmetic continuum, that of *connexity*, and that denoted by the term *perfect*, the latter is absolutely indispensable, in order that the arithmetic continuum may be suitable to be the field of operations in analysis. It will appear, when we come to the consideration of the theory of functions of a real variable, that many of the most important properties of a function may still subsist even if the domain of the variable lacks the property of connexity; but that such properties would not belong to functions of a variable which is defined for a domain such that convergent sequences of numbers in it possess no limit within that domain, and which therefore lacks the property of being perfect. This is the more remarkable on account of the fact that, in the older traditional notion of a continuum, the property of connexity was the one which was regarded as all important; the more essential property of being perfect has only been explicitly formulated in the course of the construction of the modern arithmetical theory.

42. The term arithmetic continuum is used to denote the aggregate of real numbers, because it is held that the system of numbers of this aggregate is adequate for the complete analytical representation of what is known as continuous magnitude. The theory of the arithmetic continuum has been criticized on the ground that it is an attempt to find the continuous within the domain of number, whereas number is essentially discrete. Such an objection presupposes the existence of some independent conception of the continuum, with which that of the aggregate of real numbers can be compared. At the time when the theory of the arithmetic continuum was developed, the only conception of the continuum which was extant was that of the continuum as given by intuition; but this, as we shall shew, is too vague a conception to be fitted for an object of exact mathematical thought, until its character as a pure intuitional datum has been clarified by exact definitions and axioms. The discussions connected with arithmetization have led to the construction of abstract theories* of measurable quantity; and these all involve the use of some system of arithmetic, as providing the necessary language for the description of the relations of magnitudes and quantities. It would thus appear to be highly probable that, whatever abstract conception of the intuitional continuum of quantity and magnitude may be developed, a parallel conception of the arithmetic continuum, though not necessarily identical with the one which we have discussed, will be required. To any such scheme of numbers, the same objection might be raised as has been referred to above; but if the objection were a valid one, the complete representation of continuous magnitudes by numbers would, under any theory of such magnitudes, be impossible. It is clear that it is only in connection with an exact abstract theory of magnitude that any question as to the adequacy of the continuum of real numbers for the measurement of magnitudes can arise. For actual measurement of physical, or of spatial, or temporal magnitudes, the rational numbers are sufficient; such measurement being essentially of an approximate character only, the degree of error depending upon the accuracy of the instruments employed.

---

* See O. Hölder, *Die Axiome der Quantität und die Lehre vom Mass*, Leipziger Berichte, vol. LIII (1901); also Veronese's work, *Fondamenti di Geometria*, Padua (1891); and Bettazzi's work, *Teoria delle grandezze* (1890).

The purely ordinal nature of the conception of the arithmetic continuum, including the ordinal character of an interval, has been pointed out in the course of the development of the theory. This will be further elucidated in connection with the abstract theory of order-types, to be discussed in Chapter IV.

THE CONTINUUM GIVEN BY INTUITION

43. Before the development of analysis was made to rest upon a purely arithmetical basis, it was usually considered that the field of operations was the continuum given by our intuition of extensive magnitude, especially of spatial or temporal magnitude, and of the motion of bodies through space.

The intuitive idea of continuous motion implies that, in order that a body may pass from one position A to another position B, it must pass through every intermediate position in its path. An attempt to answer the question, what is meant by *every* intermediate position, reveals the essential difficulties of this conception, and gives rise to a demand for an exact theoretical treatment of continuous magnitude.

The implication continued in the idea of continuous motion shews that, between A and B, other positions A', B' exist, which the body must occupy at definite times; that between A', B', other such positions exist, and so on. The intuitive notion of the continuum, and that of continuous motion, negate the idea that such a process of subdivision can be conceived of as having a definite termination. The view is prevalent that the intuitional notions of continuity and of continuous motion are fundamental and *sui generis*; and that they are incapable of being exhaustively described by a scheme of specification of positions. Nevertheless, the aspect of the continuum as a field of possible positions is the one which is accessible to Arithmetic Analysis, and with which alone Mathematical Analysis is directly concerned. That property of the intuitional continuum, which may be described as unlimited divisibility, is the only one that is immediately available for use in Mathematical thought; and this property is not sufficient for the purposes in view, until it has been supplemented by a system of axioms and definitions which shall suffice to provide a complete and exact description of the possible positions of points and other geometrical objects which can be determined in space. Such a scheme constitutes an abstract theory of spatial magnitude.

The exact theory of magnitude was developed to a considerable extent by Euclid; but not until recently, under the influence of the ideas of the arithmetical theory, has it been perfected in a form which exhibits the exact system of axioms and definitions necessary for a characterization of continuity that is adequate for mathematical analysis. Besides the arithmetic theory of number, there exists at the present time a theory of magnitude which runs to a certain extent parallel with the former theory. Some mathematicians* still prefer to regard number as primarily representing the ratio of two magnitudes; but they nevertheless to a large extent employ the methods of arithmetical analysis.

THE STRAIGHT LINE AS A CONTINUUM

44. Although it is no part of the plan of the present work to enter fully into the general theory of Magnitude, it is necessary briefly to consider the case of those magnitudes which are segments of a straight line; that straight line which is the ideal object of geometry, and which is the ideal counterpart of the physical straight line of perception.

The length of the segment between two points A, B, of a straight line is a particular case of a magnitude; and we shall take this conception as a datum, subject to a set of axioms† relating to the notions of congruency, and to the notions greater and less as applied to magnitudes.

We assume that any number of congruent segments OA, AB, BC, ... can be constructed on the straight line; and that any segment OA can be divided into any number of segments which are all congruent to one another.

Any segment OA may be taken as the unit of length, so that its magnitude is represented by the number 1; its multiples OB, OC, ... are denoted by the numbers 2, 3, ... If each one of the segments OA, AB, BC, ... be divided into the same number q of equal parts, then, if P is a point of division, OP is denoted by a fractional number p/q, where p is the number of the sub-segments in OP. Thus when p, q are any positive integral numbers, p/q represents a definite magnitude OP, the unit magnitude OA having been fixed upon beforehand.

Further, the number p/q may also be regarded as representing the position of the point P itself. In order to represent points of the straight line on both sides of O, the convention is made that points on one side

---

* P. Du Bois-Reymond in his *Allgemeine Functionentheorie*, Tubingen (1882), strongly advocates the view that linear magnitude forms the basis of the conception of Number. See also Stolz, *Vorlesungen uber allgemeine Arithmetik*, Leipzig (1885–86), where both views of Number are developed. See also G. Ascoli, *Rend. Ist. Lomb.* (2), XXVIII (1895).

† These axioms are discussed by O. Holder, *Leipziger Berichte*, vol. LIII (1901).

of O shall be represented by positive numbers, and those on the other side by negative numbers; thus if P is on the right of O, and P' on the left of O, and if $OP = OP'$, the point P' is represented by the number $-p/q$. The length of any segment of the straight line, whose ends are points to which rational numbers have been assigned in the manner explained above, is the difference of the above two numbers. In this manner, we have a correspondence established between the aggregate of rational numbers and an aggregate of points on the straight line, the relation of order being conserved in the correspondence, so that the two aggregates are similar.

The set of points, thus represented by rational numbers, we may speak of as the rational points of the straight line, but it must be remembered that a definite origin O, and a definite unit of length OA, are supposed to have been fixed upon beforehand; and if these be altered, the set of rational points will in general be altered also.

It has been assumed as an axiom that, if P_1Q_1 is any segment of the straight line, it may be divided into any number n_1 of equal parts: of these, if P_2Q_2 be taken as one, the same axiom asserts that P_2Q_2 may be similarly divided into any number, n_2, of equal parts, P_3Q_3 being one of the parts; and that this process may be repeated an unlimited number of times The axiom is equivalent to an assumption that the straight line is capable of unlimited divisibility; and this, being a characteristic property of the intuitional linear continuum, must also hold for its ideal counterpart, the straight line which we are here considering.

We proceed to assume as another axiom that, P_1Q_1, P_2Q_2, P_3Q_3, ... being the segments constructed as above, there exists in the straight line one point X, and one only, which separates all the points P_1, P_2, P_3, ... from all the points $Q_1, Q_2, Q_3, \ldots$. If Y be any point other than X, then points belonging to the sequence $P_1, P_2, P_3, \ldots$ and points belonging to the sequence $Q_1, Q_2, Q_3, \ldots$ can be found which are both on the same side of Y.

The point X may be regarded as the limit of either sequence of points; and the property corresponds to that property of the arithmetic continuum which is expressed by saying that it is perfect.

In accordance with this axiom there is one single point on the straight line which corresponds to any given real number; and this point, or the magnitude of the corresponding segment, may be represented by the real number.

This axiom has been stated by Dedekind, in a form corresponding to his definition of an irrational number: that a section of the rational points, in which they are divided into two classes, is made by a single point.

Another form of the axiom is that known as the *Axiom of Archimedes* that if AB, $A'B'$ are any two segments of the straight line, of which AB is the smaller one, an integer n can always be found such that $n.AB > A'B'$. As in the case of the arithmetic continuum, this is equivalent to the negation of the existence of infinitesimal segments of the straight line.

This axiom being assumed, there is a complete correspondence between the points of the straight line and the aggregate of real numbers. Thus the nature of the linear continuum, that is, so far as its possible parts, and the possible positions in it, are concerned, is completely represented and described by means of the arithmetic continuum, the axioms relating to the straight line having been so chosen that this may be the case. It will be observed that there is no real disparity between the rational points and the irrational points of the straight line; a point, which with one origin and one unit of length, is a rational point, may be an irrational point if another origin, or another unit of length, be chosen.

45. The mode which has been adopted above, of establishing a complete correspondence between the aggregate of real numbers and the aggregate of points in a straight line, though the most convenient mode, is not the only possible one. All that is really necessary for the correspondence is that, in accordance with some systematic scheme, the points in the straight line shall be made to correspond with the numbers of the arithmetic continuum, in such a way that the relation of order is conserved in the correspondence. It is not necessary that the difference of two numbers should represent the length of the segment of the straight line which is terminated by the points that correspond to the two numbers. The mode of correspondence given above is however the simplest one, and will therefore be adopted for the purpose of enabling us to use the language of geometry in analytical discussion.

In the case of space of two or of three dimensions, it will be assumed as axiomatic that one point of the space, and one only, corresponds to each pair or triplet of real numbers which represent Cartesian coordinates. This axiom may be considered as fundamental in the Cartesian system of analytical geometry.

The disputable idea that the theory here explained necessarily implies that a continuum is to be regarded as made up of points, which are elements not possessing magnitude, has frequently been a stumbling-block in the way of the acceptance of the view of the spatial continuum which has been indicated above. It has been held that, if space is to be regarded as made up of elements, these elements must themselves possess spatial character; and this view has given rise to various theories of infinitesimals

---

* The importance of the Axiom of Archimedes in this connection was pointed out and discussed by Stolz, *Math. Annalen*, vols. XXII (1883), p. 504 and XXXIX (1891), p. 107.

or of indivisibles, as components of spatial magnitude. The most modern and complete theory of this kind has been developed by Veronese*, and is based upon a denial of the principle of Archimedes which has been already referred to. In Veronese's system, when a unit segment of a straight line has been chosen, there exist segments which are too large, and others that are too small, to be capable of representation by finite numbers; and these segments are respectively infinite, and infinitesimal, relatively to the unit segment chosen. Under this scheme, a section of the rational points, or a section of the points represented by real numbers, is made, not by a single point, but by an infinitesimal segment. Veronese has consequently introduced systems of infinite and of infinitesimal numbers, each of an unlimited number of orders, for the measurement of segments which, relatively to a given scale, are infinite or infinitesimal. From his point of view, the points on a straight line which represent the real numbers form only a relative continuum, *i.e.* one which is relative to the particular scale of measurement employed; and he contemplates the conception of an absolute continuum, for the representation of which his series of sets of infinite and infinitesimal numbers are requisite. A segment, which in a given scale is finite, may be infinitesimal, or infinite of any order, when measured relatively to another scale.

The validity of Veronese's system has been criticized by Cantor and others on the ground that the definitions contained in it, relating to equality and inequality, lead to contradiction; it is however unnecessary for our purpose to enter into the controversy on this point. The straight line of geometry is an ideal object of which any properties whatever may be postulated, provided that they satisfy the conditions, (1), that they form a valid scheme, *i.e.* one which does not lead to contradiction, and (2), that the object defined is such that it is not in contradiction with empirical straightness and linearity. There is no *a priori* objection to the existence of two or more such adequate conceptual systems, each self-consistent, even if they be incompatible with one another; but of such rival schemes the simplest will naturally be chosen for actual use. Assuming then the possibility of setting up a valid non-Archimedean system for the straight line, still the simpler system, in which the principle of Archimedes is assumed, is to be preferred, because it gives a simpler conception of the nature of the straight line, and is adequate for the purposes for which it was devised. The case of the non-Euclidean systems of geometry is an instance of the existence of valid geometrical schemes divergent from one another, which nevertheless all afford a sufficient representation of physical space-percepts.

---

* See his *Fondamenti di Geometria*, Padua (1891); a German translation by Schepp has been published in Leipzig (1894).

An answer to the difficult question, in what sense the straight line, or a space of two or of three dimensions, admits of being regarded as an aggregate of points, can only be discussed after a full treatment of the nature and properties of infinite aggregates has been developed. The discussions in Chapters II, III and IV, of infinite aggregates, and especially of the notion of the *power* or *cardinal number* of such an aggregate, will throw light upon this subject.

CHAPTER II

DESCRIPTIVE PROPERTIES OF SETS OF POINTS

46. An aggregate of real numbers, each element of which consists of a single real number, is defined by any prescribed set of rules or specifications which are of such a nature that, when any real number whatever is arbitrarily assigned, they theoretically suffice to determine whether such real number does or does not belong to the aggregate. The difficulty of regarding an aggregate, so defined, as a definite object, is bound up with the difficulties connected with the notion of the linear continuum, *i.e.* the aggregate of all real numbers, out of which the defined aggregate is to be obtained by a process of selection which, except in the case of a finite aggregate, can never be actually carried out in its entirety, but which is determined by a rule or set of rules. The precise scope of the definition will be rendered clearer by the consideration of various classes of actually defined aggregates which will be considered in the present Chapter; moreover, the theoretical difficulties of the notion of such an aggregate, in general, will be in some measure elucidated by the discussions in the present and the following Chapters, of the notion of the power, or cardinal number, of an aggregate.

In accordance with the principle explained in § 44, each number of a given aggregate may be represented by a single point on a fixed straight line; thus to an aggregate of numbers there corresponds an aggregate of points on the straight line. An aggregate of single numbers, or of their equivalent points, we shall speak of as a linear set of points.

The theory of linear sets of points, of which the present Chapter contains an account, arose historically from the discussion of questions connected with the theory of Fourier's series and of the functions which can be represented by such series. A consideration of the properties and peculiarities of the sets of points at which infinities or other discontinuities of such functions exist, led to a study of the properties of linear sets in general, and to the development by G. Cantor, P. Du Bois-Reymond, Bendixson, Harnack, and others, of a general theory which has lately received wide applications both in Analysis and in Geometry.

Corresponding to the theory of linear sets of points, there exist theories of plane, solid, or p-dimensional sets of points. Each element of an aggregate in p-dimensions consists of an association of p real numbers $(x^{(1)}, x^{(2)}, \dots x^{(p)})$, and such an element is spoken of as a point in p-dimensional space. The p-dimensional continuum consists of the aggregate of all such points, when each of the numbers $x^{(1)}, x^{(2)}, \ldots x^{(p)}$ may have any value in the linear arithmetic continuum.

A considerable part of the theory of sets of points in two or more dimensions is completely parallel to the theory of linear sets of points, the proofs of many theorems for linear sets of points being applicable without essential modification to the case of sets in any number of dimensions. This remark applies both to the descriptive properties of sets considered in the present Chapter and to the metric properties which will be discussed in Chapter III. At a certain stage of the descriptive theory, however, the properties of linear sets of points exhibit a simplicity which the properties of sets in two or more dimensions do not possess. The greater complexity in the case of two or more dimensions, than in one dimension, corresponds to the greater complexity of the Analysis Situs for space of two or more dimensions than in the case of linear space.

In the present Chapter an account will be given both of those descriptive properties which are practically independent of the number of dimensions of the sets, and of those properties in which essential divergence exists between the case of linear sets and that of sets in two or more dimensions.

The whole theory of sets of points is essentially an arithmetical theory; the geometrical nomenclature and representation is a matter of convenience, not of necessity.

THE UPPER AND LOWER BOUNDARIES OF A LINEAR SET OF POINTS

47. A simple case of a linear set of points is that in which the set consists of all the points of a linear interval (a, b) either closed or open, in accordance with the definition of such an interval given in § 41.

Thus the set of points which form a closed interval (a, b) consists of all points x such that $a \leq x \leq b$; and the set of points of an open interval (a, b) consists of all points x such that $a < x < b$. Either of the sets of points x for which $a \leq x < b$, or $a < x \leq b$, may be said to be the points of a semi-closed interval (a, b), open at b or at a.

A point x of the set forming a closed interval (a, b) is said to be in the interval or segment (a, b). A point x of the set forming an open interval (a, b) is said to be within, or interior to, the interval or segment (a, b), or is said to be in the open interval (a, b).

Let a set of points be such that every point of the set lies upon a straight line, the position of each point being determined by its distance from a fixed origin upon the straight line, in the manner explained in § 44. If a point β exists, such that no number of the set is greater than β, the set is said to be bounded on the right. In this case it will be shewn that there is a definite point b, such that no point of the set is on the right of b, and such that either (1), b is itself a point of the set, or else (2), that points of the set are within the interval $(b - \epsilon, b)$, however small the positive number ϵ may be taken to be; or that both the conditions (1) and (2) are satisfied.

The point b may or may not itself be a point of the given set. In either case it is said to be the *upper boundary* of the given set. If b is itself a point of the given set, it is said to be the *upper extreme* point of the set.

When there are points of the given set interior to the interval $(b - \epsilon, b)$ for every value of ϵ $(< b)$, the point b is said to be the *upper limit* of the set.

In case b is both the upper limit, and the upper extreme point, of the set, the upper limit is said to be attained; and b may then be called the maximum point of the set.

To prove the existence*, under the condition stated, of an upper boundary, as above defined, it may be observed that all the numbers of the continuum of real numbers can be divided into two classes, one of which contains every number which is greater than all the numbers of the set, and the other of which contains every number which either belongs to the set or is less than some or all of the numbers of the set. The section thus specified defines a number b which is the upper boundary of the set.

In a similar manner, it may be shewn that, if the set is bounded on the left, *i.e.* if a point can be found such that all the points of the set are on the right of such point, then a point a exists, which is such that no points of the set are on the left of a, and such that either a is a point of the set, or else points of the set are within every interval $(a, a + \epsilon)$, where ϵ is an arbitrary positive number, or else that both conditions are satisfied simultaneously.

In case points of the set lie within every interval $(a, a + \epsilon)$, then a is called the *lower limit* of the set; and the lower limit is said to be attained if a be itself a point of the set. In any case in which a is a point of the set, it is then said to be the *lower extreme* point of the set. The term *lower boundary* may in all cases be applied to a.

A set of points which has both an upper and a lower boundary is said to be a bounded set. Thus a set is bounded if every point x in it is such that $|x| < A$, where A is some fixed positive number.

48. If no point β exists, which is such that no point of the set is on the right of β, then the set is said to be unbounded on the right; or it is said that the upper limit of the set is $+\infty$; the two statements being regarded as tautological. Similarly, if no lower boundary a exists, the set is said to be unbounded on the left; or it is said that the lower limit is $-\infty$.

---

* The existence of upper and lower boundaries was proved by Weierstrass, in his lectures. See also Bolzano, *Abh. d. Böhmischen Gesellsch. d. Wiss.*, vol. v, Prag (1817).

The symbols $+\infty$, $-\infty$ do not represent numbers of the arithmetic continuum; they must be taken to represent what is sometimes spoken of as the improperly infinite, *i.e.* the mere absence of an upper or a lower boundary respectively. In order, however, to avoid circumlocution in the statement of theorems concerning sets, it is usually convenient to speak of $+\infty$, $-\infty$, used in the above sense, as if they were numbers, sometimes called improper numbers, which correspond to upper and lower limits respectively.

The statement that $+\infty$ is the upper limit of a given set is thus taken to be equivalent to the statement that, if A is an arbitrarily chosen positive number, there exist points x of the set such that $x > A$. Similarly if $-\infty$ is the lower limit of a set, there are points x of the set such that $x < -A$.

In the present Chapter, it will frequently be assumed that the sets treated of are bounded; and the interval (a, b) will be said to be the interval in which the set exists This restriction is not so great a one as might at first sight appear, for an unbounded set can be placed into correspondence with a bounded one, in such a manner that the relative order of any two points in the one set is the same as that of the corresponding points in the other set. If $x' = \frac{x}{\sqrt{x^2+1}}$, where the radical is taken to have always the positive sign, then to a point x, in the unlimited interval $(-\infty, +\infty)$, there corresponds a point x', in the open interval $(-1, +1)$; and also $x_1' \gtreqless x_2'$, according as $x_1 \gtreqless x_2$. In order to set up a complete correspondence between the closed interval $(-1, 1)$ and the points of an unbounded segment, we must adjoin to the latter the (improper) points $+\infty$, $-\infty$, which we take to correspond respectively to the end-points $1, -1$, of the closed interval.

The same object might have been attained by using the transformation

$$x' = \frac{2}{\pi} \tan^{-1} x.$$

There is no real loss of generality in considering only such sets as lie in a given interval, say $(0, 1)$; for the relation $x' = \frac{x-\alpha}{\beta-\alpha}$ establishes a complete correspondence between sets in the interval (α, β) and sets in the interval $(0, 1)$, the relative order of points being preserved in the correspondence.

The points of the interval (α, β) may be made to correspond in order with the points of the interval $(0, 1)$, in such a manner that an arbitrarily chosen point γ within (α, β), corresponds to an arbitrarily chosen point within (0, 1); for example the point $\frac{1}{2}$. This correspondence can be effected by the transformation

$$\frac{x'}{x'-1} = \frac{x-\alpha}{x-\beta} \cdot \frac{\gamma-\beta}{\gamma-\alpha}.$$

NON-LINEAR SETS OF POINTS

49. A plane, or two-dimensional, set of points is an aggregate of which each element is constituted by a pair of real numbers $(x^{(1)}, x^{(2)})$. Such an element may be spoken of as a point P in a plane; the rectangular co-ordinates of P being $x^{(1)}, x^{(2)}$. Corresponding to intervals, in the case of linear sets of points, we now consider rectangles, or cells, of which the sides are parallel to the coordinate axes.

A closed cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ is regarded as the set of all points $(x^{(1)}, x^{(2)})$ such that $a^{(1)} \leq x^{(1)} \leq b^{(1)}$, $a^{(2)} \leq x^{(2)} \leq b^{(2)}$.

An open cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ is regarded as the set of all points $(x^{(1)}, x^{(2)})$ such that $a^{(1)} < x^{(1)} < b^{(1)}$, $a^{(2)} < x^{(2)} < b^{(2)}$.

The set of all points $(x^{(1)}, x^{(2)})$ such that $a^{(1)} \leq x^{(1)} < b^{(1)}$, $a^{(2)} \leq x^{(2)} < b^{(2)}$, may be spoken of as the semi-closed cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

A plane set of points $(x^{(1)}, x^{(2)})$ is said to be bounded, in case two positive numbers $A^{(1)}, A^{(2)}$ exist, such that $|x^{(1)}| \leq A^{(1)}$, $|x^{(2)}| \leq A^{(2)}$, for every point $(x^{(1)}, x^{(2)})$ of the set.

When all the points of a given set are in a closed cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, this cell may be spoken of as a fundamental cell for the given set of points.

More generally, a p-dimensional set of points consists of elements each of which is constituted by an association $(x^{(1)}, x^{(2)}, \ldots x^{(p)})$ of p real numbers.

The p-dimensional closed cell $(a^{(1)}, a^{(2)}, \ldots a^{(p)}; b^{(1)}, b^{(2)}, \ldots b^{(p)})$ is the set of all points $(x^{(1)}, x^{(2)}, \ldots x^{(p)})$ such that

$$a^{(1)} \leqq x^{(1)} \leqq b^{(1)},\ a^{(2)} \leqq x^{(2)} \leqq b^{(2)},\ \ldots\ldots\ a^{(p)} \leqq x^{(p)} \leqq b^{(p)}.$$

The p-dimensional open cell consists of all points such that

$$a^{(1)} < x^{(1)} < b^{(1)},\ a^{(2)} < x^{(2)} < b^{(2)},\ \ldots\ldots\ a^{(p)} < x^{(p)} < b^{(p)};$$

and the semi-closed cell consists of all points such that

$$a^{(1)} \leqq x^{(1)} < b^{(1)},\ a^{(2)} \leqq x^{(2)} < b^{(2)},\ \ldots\ldots\ a^{(p)} \leqq x^{(p)} < b^{(p)}.$$

All points of the closed cell such that $x^{(q)} = a^{(q)}$, or $x^{(q)} = b^{(q)}$ for some value of q $(= 1, 2, 3, \ldots p)$ are said to be on the boundary of the cell.

If positive numbers $A^{(1)}, A^{(2)}, \ldots A^{(p)}$ exist, such that for all points $(x^{(1)}, x^{(2)}, \ldots x^{(p)})$ belonging to a given set of points the conditions

$$|x^{(1)}| \leqq A^{(1)},\ |x^{(2)}| \leqq A^{(2)},\ \ldots\ |x^{(p)}| \leqq A^{(p)}$$

are satisfied, then the p-dimensional set is said to be bounded, and any closed cell in which all the points of the set are contained may be spoken of as a fundamental cell for the given set. If no such fundamental cell exists, for a given set of points, that set is said to be unbounded.

An unbounded set of points $(x^{(1)}, x^{(2)}, \dots x^{(p)})$ may be correlated with a bounded set $(\xi^{(1)}, \xi^{(2)}, \dots \xi^{(p)})$ for which a fundamental cell is

$$(-1, -1, \dots -1; 1, 1, \dots 1)$$

by means of the relations

$$\xi^{(q)} = x^{(q)}/\{1 + (x^{(q)})^2\}^{\frac{1}{2}}, \quad (q = 1, 2, 3, \dots p);$$

or by means of the relations

$$\xi^{(q)} = \frac{2}{\pi} \tan^{-1} x^{(q)}, \quad (q = 1, 2, 3, \dots p).$$

Any cell will be correlated with another cell by means of either of these relations. It is clear that these relations are only examples of an indefinite number of other relations by which a similar correlation may be obtained.

In order that every point on the boundary of the cell $(-1, -1, \dots; 1, 1, \dots)$ may correspond to a point in the x-space, it is convenient to adjoin a set of points at infinity to that space. For example, in the case $p = 2$, points $(x^{(1)}, \infty)$ for all values of $x^{(1)}$, points $(x^{(1)}, -\infty)$ and points $(\infty, x^{(2)})$, $(-\infty, x^{(2)})$ will be taken to correspond to the points $(\xi^{(1)}, 1)$, $(\xi^{(1)}, -1)$, $(1, \xi^{(2)})$, $(-1, \xi^{(2)})$ respectively. Also the four points $(\pm \infty, \pm \infty)$ will be adjoined so as to correspond to the four points $(\pm 1, \pm 1)$ in the finite rectangle. In this manner it can be assumed as a matter of convenience that the boundary of the rectangle $(-1, -1; 1, 1)$ corresponds to a boundary adjoined to the x-space. The advantage of this procedure is that no exception need be made as regards the correspondence of a ξ-set with an x-set, when the ξ-set has points on the boundary of the cell in which it is contained.

In stating properties of sets of points which are independent of the number of dimensions it is frequently convenient to employ a notation in which the number p of dimensions does not appear. To effect this the point $(x^{(1)}, x^{(2)}, \dots x^{(p)})$ of a p-dimensional set may be denoted by a single letter x, as in the case of a linear set. The cell $(a^{(1)}, a^{(2)}, \dots a^{(p)}; b^{(1)}, b^{(2)}, \dots b^{(p)})$ may then be denoted by (a, b). Thus (a, b) denotes an interval when $p = 1$, and a cell when $p > 1$.

LIMITING POINT OF A CONVERGENT SEQUENCE OF INTERVALS, OR CELLS

50. Let (a_1, b_1), (a_2, b_2), ... (a_n, b_n), ... be an unending sequence of closed intervals which are such that any one of them (a_n, b_n) lies entirely in the preceding one (a_{n-1}, b_{n-1}), the two having at most one end-point common; thus $a_n \geqq a_{n-1}$, $b_n \leqq b_{n-1}$; moreover, suppose that the lengths $b_1 - a_1$, $b_2 - a_2$, ... $b_n - a_n$, ... form a sequence which converges to zero, the condition for which is that, corresponding to any arbitrarily small ϵ, n can be so chosen that $b_m - a_m$, for all values of m which are $\geqq n$, is $< \epsilon$.

It will be seen that, in accordance with the axioms explained in § 14, *there exists one point and one only which is in every interval of the sequence.* This point may be called the *limiting point* of the convergent sequence of closed intervals.

Each of the aggregates $(a_1, a_2, a_3, \dots a_n, \dots)$, $(b_1, b_2, \dots b_n, \dots)$ being convergent, defines a number; and in fact, in virtue of the definition of equality in § 26, they define the same number x. This number x is not less than a_n and not greater than b_n, whatever n may be, the point x therefore lies in all the intervals, and is the limiting point whose existence was to be remarked If y be any number greater (or less) than x, we can find n so great that $b_n - x < y - x$, if $y > x$, or that $x - a_n < x - y$, if $x > y$; thus y does not lie in (a_n, b_n). Hence there is only one point which satisfies the prescribed conditions.

If, for every n, from and after some fixed value, the inequalities $a_n > a_{n-1}$, $b_n < b_{n-1}$ both hold, then the limiting point x is in the interior of all the intervals of the sequence. If, from and after some fixed value of n, say n_1, we have $a_n = a_{n-1}$, $b_n < b_{n-1}$, the limiting point x coincides with the common end-points $a_{n_1-1}, a_{n_1}, a_{n_1+1}, \dots$.

If the intervals (a_1, b_1), (a_2, b_2), $\dots (a_n, b_n)$, $\dots$ are all open intervals, there is not necessarily any point which is in all the open intervals. Consider, for example, the set of open intervals

$$(0, 1), \left(0, \frac{1}{2}\right), \left(0, \frac{1}{3}\right), \dots \left(0, \frac{1}{n}\right), \dots.$$

In this case the point x defined by either of the sequences

$$(0, 0, 0, \dots), \left(1, \frac{1}{2}, \frac{1}{3}, \dots \frac{1}{n}, \dots\right)$$

is the point 0, which is not in any of the open intervals of the given set. It thus appears that the theorem does not hold for the set of open intervals if, from and after some fixed value of n, we have $a_n = a_{n+1}$, or if, from and after some fixed value of n, we have $b_n = b_{n+1}$. If however, from and after some fixed value of n, both the conditions $a_n < a_{n+1}$, $b_n > b_{n+1}$ are satisfied, the theorem is valid; for the point x, defined as above, is a point in every open interval (a_n, b_n), from and after some fixed value of n.

If the intervals of the given set are semi-closed; being all open at the right-hand end-points, the theorem holds good unless $b_n = b_{n+1}$, from and after some fixed value of n.

In the case of a sequence of p-dimensional closed cells

$$(a_n^{(1)}, a_n^{(2)}, \dots a_n^{(p)};\ b_n^{(1)}, b_n^{(2)}, \dots b_n^{(p)}),$$

where $n = 1, 2, 3, \dots$, each of which is in the preceding one, so that $a_{n+1}^{(q)} \geq a_n^{(q)}$, and $b_{n+1}^{(q)} \leq b_n^{(q)}$, for all values of n, and for the values $1, 2, 3, \dots, p$, of q, and if $b_n^{(q)} - a_n^{(q)}$, for each value of q, forms a sequence of numbers which converges to zero, for $n = 1, 2, 3, \ldots$, there is a unique point which is in all the cells of the set, and which is called the limiting point of the set. For the sequence of linear intervals $(a_n^{(q)}, b_n^{(q)})$ defines, for each value of q, a single point $x^{(q)}$ in all the intervals. Thus the point $(x^{(1)}, x^{(2)}, \ldots x^{(p)})$, uniquely determined, is in all the cells of the sequence, and it is easily seen that it is the only point that satisfies this condition.

If the cells are open, the corresponding theorem does not hold, unless there exists some value of n from and after which $a_n^{(q)} < a_{n+1}^{(q)}$, $b_n^{(q)} > b_{n+1}^{(q)}$, for each value of q.

If the cells are semi-closed, for the validity of the theorem the condition $b_n^{(q)} > b_{n+1}^{(q)}$ must be satisfied, from and after some fixed value of n, for every value of q.

The number $\left\{\sum_{q=1}^{q=p} (b^{(q)} - a^{(q)})^2\right\}^{\frac{1}{2}}$ may be called the *span*, or the *length of the diagonal*, of the cell $(a^{(1)}, a^{(2)}, \ldots a^{(p)}; b^{(1)}, b^{(2)}, \ldots b^{(p)})$. This number may also be spoken of as the distance between the two points

$$(a^{(1)}, a^{(2)}, \ldots a^{(p)});\ (b^{(1)}, b^{(2)}, \ldots b^{(p)}),$$

and may be denoted by PQ, when the points are denoted by P, Q respectively.

A convergent sequence of closed cells is one in which each cell is in the preceding one, and such that the sequence of spans of the cells converges to zero; such a convergent sequence defines a unique point in all the cells of the sequence.

In the case of such a convergent sequence of cells, the convergence of the sequence of spans to the limit zero ensures that, for each value of q, where $1 \leq q \leq p$, $b_n^{(q)} - a_n^{(q)}$ converges to zero, as n is indefinitely increased. A sequence of closed cells, each of which contains the next, may however be such that $b_n^{(q)} - a_n^{(q)}$ converges to zero for some of the values of q, but not for all these values. In that case the sequence of cells may be said to be ***incompletely convergent***; the points of a cell of dimensions lower than p will all be contained in each of the cells of the given sequence. This cell may be termed the ***limiting cell*** (or interval) of the given sequence.

For example, if $p = 3$, we may have $b_n^{(1)} - a_n^{(1)}$ and $b_n^{(2)} - a_n^{(2)}$ converging to zero, but $b_n^{(3)} - a_n^{(3)}$ converging to some number greater than zero; in that case the sequence converges to a limiting interval. Again, if $b_n^{(1)} - a_n^{(1)}$ converges to zero, but $b_n^{(2)} - a_n^{(2)}$, $b_n^{(3)} - a_n^{(3)}$ converge to numbers greater than zero, the limiting cell is two-dimensional.

SYSTEMS OF NETS

51. Let the linear interval (a, b) be divided into a number m_1 of parts, such that the length of each part does not exceed a positive number δ_1; let D_1 denote the finite set of intervals so obtained. By introducing further points of division of (a, b), let a new finite set of intervals D_2 be determined, of which the number is m_2 $(> m_1)$, and let the length of each interval of D_2 not exceed a positive number δ_2 $(< \delta_1)$. Proceeding in this manner, there can be defined, in accordance with some specified set of rules, a sequence $D_1, D_2, \dots D_n, \dots$ of finite sets of intervals, such that the intervals of D_n are all not greater than an assigned number δ_n. Let it be supposed that the numbers $\delta_1, \delta_2, \dots \delta_n, \dots$ form a diminishing sequence which converges to zero.

If we regard the intervals of each set D_n as semi-closed, *i.e.* closed on the left and open on the right, except that the extreme interval on the right is taken to be closed, each point of the closed interval is in one and only one of the intervals of the set D_n, for each value of n.

The set D_n of semi-closed intervals (except that the extreme one on the right is closed) may be spoken of as a *net*, and the nets corresponding to $n = 1, 2, 3, \dots$ may be called *a system of nets fitted on to the interval* (a, b). The particular net D_n may be called the net of order n, or the nth net of the system. The m_n separate intervals of the net D_n may be spoken of as the *meshes* of the net D_n.

The fundamental property of such a system of nets is that each point x of the interval (a, b) is in one and only one mesh of each of the nets D_n. These meshes, for $n = 1, 2, 3, \dots$, constitute a convergent sequence of intervals of which the point x is the limiting point. In each net of the system the rank, or order, of the meshes may be taken as their order from left to right in the interval (a, b).

In case D_1 consists of m meshes of equal breadth, D_2 of m^2 meshes of equal breadth, and in general D_n consists of m^n meshes of equal breadth, the system of nets will be said to be *symmetrical*.

In some cases it is convenient to employ nets in which all the meshes are closed at both ends. In that case a point x of (a, b) that is an end-point of a mesh of D_n will, in general, belong to two adjacent meshes of D_n, and of all the subsequent nets of the system. Such a point will consequently be the limiting point of more than one sequence of meshes. Such a system of nets may be spoken of as *a system of nets with closed meshes*. It may or may not be a symmetrical system.

In case the interval (a, b) is replaced by an indefinite interval $(-\infty, \infty)$, (a, ∞), or $(-\infty, b)$, the number of meshes of a net fitted on to such indefinite interval will not be finite, but as before, the set of numbers $\delta_1, \delta_2, \dots \delta_n, \dots$ will be taken to converge to zero.

The definition of a system of nets can be extended to the case of cells of two, or of any number p, dimensions. In the case of the two-dimensional cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, this rectangle is divided by straight lines parallel to the axes into a set D_1 of rectangular cells, the spans of all of which are $\leq \delta_1$. Some or all of these cells are again divided into smaller cells, by the introduction of fresh straight lines parallel to the axes, so that the new set of cells D_2 is such that their spans are all $\leq \delta_2$. Proceeding in this manner, in accordance with a specified set of rules, and the diminishing sequence $\{\delta_n\}$ being taken to converge to zero, we obtain a sequence $D_1, D_2, \dots D_n, \dots$ of sets of cells. We take each of the cells of D_n to be semi-closed, except that those which have a side in common with one of those sides of $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ which do not contain the point $(a^{(1)}, a^{(2)})$ are closed along that side; that side itself being taken to be semi-closed, unless it has $(a^{(2)}, b^{(2)})$ for an end-point, in which case it is regarded as closed. Each of the sets D_n of rectangular cells is then termed a net; each of the cells of which D_n is composed is termed a mesh of the net D_n, and the totality of the nets $\{D_n\}$ is called a system of nets fitted on to the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

Any point x in the fundamental cell is now in one and one only of the meshes of D_n, for each value of n; and consequently x is the limiting point of a unique sequence of meshes of the cells of the system.

It is easy to arrange the meshes of a net D_n in ascending order or rank. Thus we may take those cells which have a side in common with a part of $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ in ascending order from left to right; after these we may take in order from left to right those cells which have a side in common with one of those just referred to; then we take in order from left to right those cells which have a side in common with one of these latter, and so on.

If the nets of D_n all have the same span δ_n for each value of n, and if the ratio of δ_{n+1} to δ_n is independent of n, the system of nets is said to be symmetrical.

As before, a system of nets may be fitted on to the unbounded plane of $(x^{(1)}, x^{(2)})$.

In case all the meshes of D_n, for every value of n, are closed, the system is said to be a system of nets with closed meshes. In this case the sequence of meshes of which the limit is a prescribed point $(x^{(1)}, x^{(2)})$ is in general not unique.

It is unnecessary to point out in detail the corresponding definition and convention for the case of a p-dimensional cell, when $p > 2$. It is clear that the number of dimensions makes no difference as regards the properties of a system of nets.

It will appear that a system of nets provides an apparatus which is of considerable utility as an instrument of investigation in the theory of sets of points, and in that of functions of one or more variables. The employment of this nomenclature has the advantages of preventing repetition in the proofs of various important theorems, and of providing the means of giving those proofs in a form applicable not only to linear sets of points or to functions of one variable, but also to sets of any number of dimensions, or to functions of any number of variables.

Although the method, equivalent to the employment of systems of nets, has been in use for a considerable time past, the first formulation of the idea of a system of nets appears to be due to de la Vallée Poussin*. His terminology is however not identical with that here employed. He speaks of what is here termed a system of nets, as simply a net (réseau), and what is here called a net is termed by him a lattice (grillage). Moreover he considers only what are here described as symmetrical systems of nets, and this is sufficient for the special purpose for which he employs them. For the more general purposes of investigation, it is however convenient not to make this restriction, accordingly the wider definition given above has been here adopted.

THE LIMITING POINTS AND THE DERIVATIVES OF A SET

52. If a point x be taken in the interval (a, b), an interval $(x - \epsilon_1, x + \epsilon_2)$ which lies entirely in (a, b) is called a *neighbourhood* of the point x; and this neighbourhood may be made as small as we please by proper choice of ϵ_1 and ϵ_2. An interval $(x, x + \epsilon_2)$ is called a neighbourhood of x on the right, and $(x - \epsilon_1, x)$ is called a neighbourhood of x on the left. The end-points a and b can only have neighbourhoods on the right and the left respectively. A neighbourhood of a point may be open or closed.

In the case of a point $(x^{(1)}, x^{(2)})$ in a plane, the cell

$$(x^{(1)} - \epsilon_1^{(1)}, x^{(2)} - \epsilon_1^{(2)};\ x^{(1)} + \epsilon_2^{(1)}, x^{(2)} + \epsilon_2^{(2)})$$

is called a neighbourhood of $(x^{(1)}, x^{(2)})$; it may be closed or open. Corresponding to neighbourhoods on the right or left, of a point x of a linear interval, there are in the plane four partial neighbourhoods of the point $(x^{(1)}, x^{(2)})$, viz. the four rectangles

$$(x^{(1)}, x^{(2)};\ x^{(1)} + \epsilon_2^{(1)}, x^{(2)} + \epsilon_2^{(2)}),\ (x^{(1)} - \epsilon_1^{(1)}, x^{(2)} - \epsilon_1^{(2)};\ x^{(1)}, x^{(2)}),$$
$$(x^{(1)} - \epsilon_1^{(1)}, x^{(2)};\ x^{(1)}, x^{(2)} + \epsilon_2^{(2)}),\ (x^{(1)}, x^{(2)} - \epsilon_1^{(2)};\ x^{(1)} + \epsilon_2^{(1)}, x^{(2)}).$$

In p dimensions, similar definitions hold good; a point has 2^p partial neighbourhoods. The neighbourhoods can always be supposed to be in a fundamental cell, to which the point is confined.

If a set of points G, in any number of dimensions, has been defined, and a point P be such that every neighbourhood of it contains a point of G, other than P itself, the point P is called a limiting point of the set G, whether P belongs to G or not.

---

* See his treatise *Intégrales de Lebesgue*, Paris (1916), p. 16, *et seq.*

It should be observed that, if every neighbourhood of P contains a point of G, other than P, it must contain an infinite number of points of G. For let us suppose that a certain neighbourhood H, of P, contains only a finite set of points $p_1, p_2, \dots p_m$ belonging to G. We may suppose P to be the limiting point of the sequence $\{d_n\}$ of meshes belonging to the nets $\{D_n\}$ of a system of nets fitted on to H, and such that P is interior to all the meshes d_n. Each of the points $p_1, p_2, \dots p_m$ can be in only a finite number of the meshes d_n of the sequence; hence, from and after some fixed value of n, d_n contains no point of G; but d_n is a neighbourhood of P containing no points of G, other than P itself, which is contrary to the hypothesis made above.

A limiting point of G, defined as above, is also called a *point of accumulation*. Another definition of a limiting point is that it is a point P such that G contains a sequence of points $\{P_n\}$ for which the distance PP_n is less than an arbitrarily chosen number ϵ, for all sufficiently large values of n. A point which satisfies this condition is certainly a limiting point. in accordance with the definition here adopted, but the question whether the converse holds will be discussed in Chap. IV, in connection with the multiplicative axiom.

The fundamental theorem will now be established, that: *Every bounded set of points G which contains an infinite number of points possesses at least one limiting point*

We may suppose a system of nets to be fitted on to the fundamental cell, or interval, in which G is contained. Of the meshes of D_1 at least one must contain an infinite set of points belonging to G; let d_1 be that one of such meshes which is of lowest rank. Of the meshes of D_2 which are contained in d_1 at least one must contain an infinite set of points belonging to G; let d_2 be that one of such meshes which is of lowest rank. Proceeding in this manner a definite sequence of meshes $d_1, d_2, \dots d_n, \dots$ is defined, each of which contains an infinite set of points belonging to G. The limiting point P of this sequence $\{d_n\}$ of meshes is a limiting point of G. For any assigned neighbourhood of G contains d_n, from and after some fixed value of n, and therefore contains an infinite number of points of G.

It has thus been shewn that G has at least one limiting point. It may have a finite number, or an indefinitely great number, of limiting points. It should be observed that a limiting point of G may or may not itself be a point of G. In the case of a linear set of points, if either boundary of the set be not a point of G, it is certainly a limiting point of G; it may however be both a point of G and a limiting point of G.

A limiting point P of a linear set G is a limiting point on both sides, if an indefinitely great number of points of G lie in every neighbourhood of P on the right, and also in every neighbourhood of P on the left.

Otherwise P is a limiting point of G on one side only. This distinction may be extended to the case of a limiting point P of a set G in p dimensions, the 2^p partial neighbourhoods of P corresponding to neighbourhoods on the right and left of a point of a linear set.

53. The theorem of § 52 does not hold for the case of unbounded sets of points. For example, the linear set which consists of the points $1, 2, 3, \ldots, n, \ldots$ has no limiting point, in accordance with the definition given in § 52. The same remark applies to the set of points $-1, -2, -3, \ldots, -n, \ldots$. If we apply to either of these sets the transformation

$$\xi = x/\sqrt{x^2+1},$$

where the positive value of the radical is always taken, the sets on the interval of ξ which correspond to the above sets are

$$1/\sqrt{2}, \quad 2/\sqrt{5}, \quad 3/\sqrt{10}, \quad n/\sqrt{n^2+1}, \quad \ldots;$$

and

$$-1/\sqrt{2}, \quad -2/\sqrt{5}, \quad -3/\sqrt{10}, \quad -n/\sqrt{n^2+1}, \quad \ldots,$$

which have as limiting points the points $1, -1$ respectively; these points $1, -1$ being the end-points of the interval $(-1, 1)$ of ξ. To the indefinitely great interval of x, which is an open interval, there corresponds the open interval $(-1, 1)$ of ξ, and in the former interval there are no points which correspond to the points $-1, 1$ of the interval of x. If we agree to adjoin to the set of real numbers, two improper numbers $-\infty, \infty$, which are taken to correspond to the numbers $-1, 1$ in the closed interval $(-1, 1)$ of ξ, we now regard $(-\infty, \infty)$ as the closed interval of x corresponding to the closed interval $(-1, 1)$ of ξ. To a neighbourhood $(1-\epsilon, 1)$ on the left of the point $\xi = 1$, there will correspond the interval

$$\left(\frac{1-\epsilon}{(2\epsilon-\epsilon^2)^{\frac{1}{2}}}, \infty\right)$$

on the left of the adjoined point $x = \infty$ in the linear interval of x. The point $x = \infty$ may be regarded, in an extended sense of the term, as the limiting point of the set of points $1, 2, 3, \ldots$. It corresponds to the limiting point 1 of the corresponding set $1/\sqrt{2}, 2/\sqrt{5}, 3/\sqrt{10}, \ldots\ldots$ in the segment $(-1, 1)$.

In this extended sense of the term "limiting point," the point ∞ will be a limiting point of any set G which is such that, corresponding to any arbitrarily chosen positive number A, there are points x, of G, for which $x > A$. Similarly, the point $-\infty$ may be regarded as a limiting point of the set G, if there are points of G for which $x < -A$, for all values of the positive number A. When ∞ or $-\infty$, is, in this extended sense of the term, a limiting point of G, the point 1, or -1, is a limiting point of the set of points which corresponds to G in the ξ-segment.

To a neighbourhood $(x - \epsilon_1, x + \epsilon_2)$ of a point x, not ∞ or $-\infty$, in the interval $(-\infty, \infty)$, there corresponds a neighbourhood $(\xi - \epsilon_1', \xi + \epsilon_2')$ of the point ξ which corresponds to x. It is easily seen that sequences of values of $\epsilon_1 . \epsilon_2$ which converge to zero, correspond to sequences of ϵ_1', ϵ_2' which converge to zero. Thus a finite limiting point of a set G in the interval $(-\infty, \infty)$ of x corresponds to a limiting point of the corresponding set in $(-1, 1)$ which is in the interval.

It thus appears that, when the two points $+\infty$, $-\infty$ are adjoined to the indefinite interval of x, so that it becomes a closed interval, the theorem as to the existence of a limiting point holds of the closed interval. The difference in the form of the condition that either of the points ∞, $-\infty$ should be a limiting point, from the condition applicable to a finite point, is seen to be unessential, as it disappears when the set is transformed into a set in the finite closed interval $(-1, 1)$.

The intervals (A, ∞), $(-\infty, -A)$ may be termed neighbourhoods of the points ∞, $-\infty$ respectively.

The case of a plane set may be considered in a similar manner. Employing the transformation

$$\xi^{(1)} = x^{(1)}/\sqrt{(x^{(1)})^2 + 1}, \quad \xi^{(2)} = x^{(2)}/\sqrt{(x^{(2)})^2 + 1},$$

by which points $(x^{(1)}, x^{(2)})$ of the unclosed infinite cell $(-\infty, -\infty, \infty, \infty)$ are placed in correspondence with points of the unclosed cell $(-1, -1; 1, 1)$ in the plane of $(\xi^{(1)}, \xi^{(2)})$; we see that, corresponding to a set of points in the latter plane which has a limiting point on the boundary of the rectangle $(-1, -1; 1, 1)$, there will be a set of points in the plane of $(x^{(1)}, x^{(2)})$ which has a limiting point, in the extended sense of the term, on a boundary adjoined to the rectangle $(-\infty, -\infty; \infty, \infty)$. In this case also, any limiting point of a set in the rectangle $(-1, -1, 1, 1)$ in which the points $(\xi^{(1)}, \xi^{(2)})$ are contained, will correspond to a limiting point of the corresponding set in the rectangle $(-\infty, -\infty; \infty, \infty)$ in which the points $(x^{(1)}, x^{(2)})$ are contained. It is clear that there is an arbitrary element in the particular transformations employed.

The set of points $(x^{(1)}, x^{(2)})$ such that $a^{(1)} - \epsilon \leqq x^{(1)} \leqq a^{(1)} + \epsilon'$, and $x^{(2)} \geqq A$, where ϵ, ϵ', A are positive numbers, may be called a closed neighbourhood of the point $(a^{(1)}, \infty)$. A similar definition may be given for an open neighbourhood of the point. The set of points for which $x^{(1)} \geqq A$, $x^{(2)} \leqq -B$, where A and B are positive numbers, may be termed a closed neighbourhood of the point $(\infty, -\infty)$; if $x^{(1)} > A$, $x^{(2)} < -B$, the neighbourhood is open. Similar definitions will apply to points at infinity of the other types.

54. Returning to the case of a set G in a finite interval or cell, we observe that the limiting points of G form a set of points which may be finite or infinite; this set is called the *derived set**, or *first derivative* of G, and may be denoted by G'. In case the set G' contains an infinite number of points, it possesses itself a derivative set G'', which is called the *second derivative* of G. If we proceed in this manner, we may obtain a series

$$G', G'', G''', \dots G^{(n)}$$

of derivatives of G. If the nth derivative $G^{(n)}$ contains a finite number only of points, then these have no limiting point, and we may say that $G^{(n+1)} \equiv 0$. It may however happen that, however large the integer n may be, the derivative $G^{(n)}$ contains an indefinitely great number of points; and thus a next derivative exists.

A set G which possesses only a finite number of derivatives is said to be of the first species.

In this case, if $G^{(s)}$ contains only a finite number of points, the set G is said to be of order† s. Thus, for example, a set of the first species and order zero contains only a finite number of points; and a set of the first species and order 1 has a first derivative which contains only a finite number of points. It will be observed that the order of each derivative of G is less by unity than that of the one which precedes it.

A set G which possesses an indefinite number of derivatives is said to be of the second species.

As an example, we may consider the set of rational numbers in the interval (0, 1). The first derivative of this set contains every real number in (0, 1), and all subsequent derivatives are identical with the first.

EXAMPLES

1. Let‡ $$G=\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4} \dots \frac{1}{n} \dots\right).$$

We see that G' consists of the single point 0, which does not belong to G; thus G is of the first species and of order 1.

2. Let§ the points of G be given by

$$\frac{1}{3^{s_1}}+\frac{1}{5^{s_2}}+\frac{1}{7^{s_3}}+\frac{1}{11^{s_4}},$$

where s_1, s_2, s_3, s_4 each have all positive integral values. Here G' consists of the four sets of points given by

$$\frac{1}{3^{s_1}}+\frac{1}{5^{s_2}}+\frac{1}{7^{s_3}},\quad \frac{1}{3^{s_1}}+\frac{1}{5^{s_2}}+\frac{1}{11^{s_4}},\quad \frac{1}{3^{s_1}}+\frac{1}{7^{s_3}}+\frac{1}{11^{s_4}},\quad \frac{1}{5^{s_2}}+\frac{1}{7^{s_3}}+\frac{1}{11^{s_4}};$$

---

* The notion of the derivative of a set was introduced by Cantor, *Math. Annalen*, vol. v (1872), p. 128. Du Bois-Reymond contemplated the existence of limiting points of various orders, *Crelle's Journal*, vol. LXXIX (1874), p. 30; in *Math. Annalen*, vol. XVI (1880), p. 128, Du Bois-Reymond defined a limiting point of infinite order.

† Cantor, *Math. Annalen*, vol. v (1872), p. 129.

‡ Cantor, *Math. Annalen*, vol. v (1872).

§ Ascoli, *Ann. di Mat.* (2), vol. VI (1875), p. 56.

and of the six sets of points

$$\frac{1}{3^{s_1}}+\frac{1}{5^{s_2}},\quad \frac{1}{3^{s_1}}+\frac{1}{7^{s_3}},\quad \frac{1}{3^{s_1}}+\frac{1}{11^{s_1}},\quad \frac{1}{5^{s_2}}+\frac{1}{7^{s_3}},\quad \frac{1}{5^{s_2}}+\frac{1}{11^{s_4}},\quad \frac{1}{7^{s_3}}+\frac{1}{11^{s_4}};$$

and of the four sets of points $\frac{1}{3^{s_1}},\ \frac{1}{5^{s_2}},\ \frac{1}{7^{s_3}},\ \frac{1}{11^{s_4}};$

together with the single point 0. G'' consists of the last ten of these sets, and of the point 0. The third derivative G''' consists of the last four sets, and of the point 0; G'''' consists of the point 0 only. The set G is of the first species and of the fourth order.

3. Let* the points of G be given by

$$\frac{1}{a_1}+\frac{1}{a_2}+\ldots+\frac{1}{a_n},$$

where n is a fixed number, and each of the numbers $a_1, a_2, \ldots a_n$ takes every positive integral value. In this case G is of order n.

4. The zeros† of the function $\sin\frac{1}{x}$ form a set similar to that in Example 1. The zeros of the function $\sin\left(\frac{1}{\sin\frac{1}{x}}\right)$ form a set of the second order, those of $\sin\left(\frac{1}{\sin\frac{1}{\sin\frac{1}{x}}}\right)$ form a set of the third order, and so on.

5. Let‡ the points of G be given by

$$\frac{1}{2^{m_1}}+\frac{1}{2^{m_1+m_2}}+\ldots+\frac{1}{2^{m_1+m_2+\ldots+m_n}},$$

where $m_1, m_2, \ldots m_n$ have all positive integral values, including zero, and n is a fixed integer. It can be seen that $G^{(n)}$ consists of the point zero only.

DESCRIPTIVE TERMINOLOGY

55. If $G_1, G_2, G_3, \ldots G_n$ denote a number of sets of points (either linear or in any number of dimensions), the set which contains every point that belongs to one or more of the given sets is called§ their *least common multiple*, and is denoted by $M(G_1, G_2, \ldots G_n)$. In case no two of the given sets have a point in common, the common measure of the sets may be denoted by $G_1 + G_2 + \ldots + G_n$, and it may be spoken of as their sum. By some writers the term "sum" is employed for the greatest common measure.

That set which contains all those points which belong to every one of the given sets is called their *greatest common divisor*, and it may be denoted by $D(G_1, G_2, \ldots G_n)$. By some writers this set is termed the product of the given sets, and is denoted by $G_1G_2 \ldots G_n$.

These definitions are still applicable in the case of an unending sequence of sets $G_1, G_2, G_3, \ldots$.

---

* H. J. S. Smith, *Proc. Lond. Math. Soc.* vol. VI (1875), p. 145.

† P Du Bois-Reymond, *Crelle's Journal*, vol. LXXIX (1875), p 36.

‡ Mittag-Leffler, *Acta Math.* vol. IV (1884), p. 58.

§ Cantor, *Math. Annalen*, vol. XVII (1880), p. 355.

If all the points of a set H are points of a set G, H is said to be contained in G, or to be a *part*, or *component*, of G. The set G is said to contain H. Those points of G that do not belong to H form a set which may be denoted by $G - H$. The set $G - H$ is said to be the *complement* of H with respect to G, and is sometimes denoted by $C_G(H)$. If the set G consists of all the points of the fundamental interval, or cell (open or closed), which contains H, the set $G - H$ is called the complement of H, and is denoted by $C(H)$.

A set, all of whose limiting points belong to the set itself, is said to be closed.

Thus, in a closed set, the derivative G' is a component of G.

Although the definition of a closed set would be applicable to an unbounded set, in case limiting points in the extended sense of the term are admitted, as explained in § 53, it is usually convenient to restrict statements as to closed sets to the case in which they are bounded. Accordingly, it will in general be assumed that a closed set is a bounded one

A set of points G is said to be an isolated set when no point of the set is a limiting point of the set.

Thus, we have, for an isolated set G, the condition $D(G, G') \equiv 0$.

If, from any set G, we remove those points which also belong to its derivative, the remainder is an isolated set; thus $G - D(G, G')$ forms an isolated set. Any set G may be regarded as the sum of an isolated set and of a component of the derivative G'.

If a component H of the set G is such that every point of G is a limiting point of H, the set H is said to be dense in G.

If we consider the case in which H is identical with G, we obtain the definition:

If every point of G is a limiting point of the set, G is said to be dense in itself.

For a set G, dense in itself, G is a component of the derivative G'. The rational numbers of the interval (0, 1) form a set that is dense in itself.

*A set G which is both closed and dense in itself is said to be perfect**.

Thus a perfect set G is identical with its derivative. It follows that every perfect set is of the second species.

By some writers† the term *perfect* is applied to sets which, in accordance with the terminology of Cantor here adopted, are only closed, without necessarily being dense in themselves; what is here called a perfect set is then spoken of as an *absolutely*‡ *perfect set.*

---

* Cantor, *Math. Annalen*, vol. XXI (1883), p. 575.

† For example Jordan, see *Cours d'Analyse*, vol. I, Paris (1893), p. 19.

‡ Borel, *Leçons sur la théorie des fonctions* (1898), p. 36.

In case the set G consists of all the points of a closed interval, or cell, a set H which is dense in G is said to be everywhere dense[*] in the given interval, or cell. This is equivalent to the statement that:

A set H is said to be everywhere dense, or dense in an interval, or in a cell, provided that no interval, or cell, contained in the given one, exists which contains no points of H.

Similarly, if a set H is dense in a set G, no interval, or cell, can be determined such as to contain points of G and no points of H.

If H is dense in G, the derivative H', of H, contains every point of G. The derivative G' of a set G which is everywhere dense in an interval, or cell, must contain every point of the closed interval, or cell. This property may be used as a definition[†] of the term "everywhere dense."

By Du Bois-Reymond[‡] the term *pantachisch* was used with the same meaning as everywhere dense.

Any interval, or cell, contained in a fundamental interval, or cell, may be spoken of as a *sub-interval*, or *sub-cell*

If, in every sub-interval, or sub-cell (a', b'), of the fundamental interval, or cell (a, b), in which a set of points G is contained, another sub-interval, or sub-cell (a'', b''), can be determined which contains no points of G, the set G is said to be nowhere dense, or non-dense in (a, b)

Thus a non-dense set is one such that no interval or cell exists in which the set is everywhere dense.

A component H of a set G is said to be non-dense in, or relatively to G, if, in any interval, or cell, that contains points of G, a sub-interval, or sub-cell, can be determined which contains a point of G but no point of H.

A point P of a set G is said to be an interior point of G, if a neighbourhood of P can be determined all the points of which are points of G.

If, however, the set G be bounded, and contained in a fundamental interval, or cell, a point of G on the boundary of the interval, or cell, may be regarded as an interior point of G relatively to the interval or cell, if a neighbourhood of the point exists such that every point of that neighbourhood which is in the fundamental interval, or cell, is a point of G. Such a point is however not an interior point of G, relatively to unbounded space, or to an interval, or cell, which contains the fundamental interval or cell in its interior.

Those points of a set G which are not interior points of G, together with those points of $C(G)$ which are not interior points of $C(G)$, form a set which

---

[*] Cantor, *Math. Annalen*, vol. xv (1879), p. 2.

[†] Baire, *Annali d. Mat.* (3), vol. III (1899), p. 29.

[‡] *Math. Annalen*, vol. xv (1879), p. 287.

is called the frontier of G or of C (G); the set C (G) denoting the complement of G in the unbounded interval or space in which G is contained.

If, however, G is contained in an interval, or cell, and C' (G) denotes the complement of G with respect to such interval, or cell, the set of points of G not interior to G, relatively to the interval or cell, together with those points of C' (G) which are not interior to C (G), relatively to the interval or cell, forms the boundary of G or of C (G) relatively to the fundamental interval, or cell.

To illustrate the distinction here made, we consider for example the linear set consisting of all the points of the closed interval (a, b); the boundary of G consists of the two points a, b, which are not interior points either of G or of C (G), the complement of G relatively to $(-\infty, \infty)$. But, relatively to the closed interval (a, b), the points a, b are interior points of the closed set (a, b), in accordance with the convention made above; and thus the set has no boundary relatively to the closed interval (a, b).

Every point of an open interval, or cell, is an interior point of the open interval, or cell, regarded as a set of points. For example, every point of an unbounded space is an interior point of the unbounded space considered as a set of points.

An open set is one in which every point is an interior point. A set is open relatively to a fundamental interval, or cell, when every point is an interior point relatively to the interval, or cell.*

The term open set is however employed by some writers to denote any set which is not closed.

A non-dense set has no interior points, but an everywhere dense set may also have no interior points. An open set contains none of the points of its frontier.

It is frequently of importance to consider the properties of sets which are contained in a given perfect set G, or which have a part in common with G.

A point P of a set H is said to be an interior point of H relatively to G, if it is a point of G and is such that a neighbourhood of P exists for which all the points of G in that neighbourhood are also points of H.

Those points of G which are limiting points both of H and of C_G (H), where H is a component of G, are said to form the frontier of H and of C_G (H) relatively to G.

A set H, all the points of which are interior points of H relatively to G, is said to be open relatively to G.

A set H, such that no point which it has in common with G is an interior point of H relatively to G, is said to be diffuse relatively to G, or to be diffused in G.

---

* de la Vallée Poussin, *Intégrales de Lebesgue*, p. 10.

If H has points in common with G, and is not diffuse in G, it is said to be compact in G.*

This property of compactness was formulated by de la Vallée Poussin, whose definition is equivalent to that here given. The term "compact" is employed by some writers with a different meaning.

PROPERTIES OF CLOSED AND OPEN SETS

56. *The complement $C(G)$ of a closed set G with respect to a closed interval, or cell, in which G is contained, is an open set, relatively to the interval or cell. Conversely, the complement of an open set contained in a closed interval, or cell, is a closed set.*

For any point P of $C(G)$ which is not a limiting point of the closed set G is such that a neighbourhood of P can be determined which contains no points of G. All the points of such neighbourhood (or of the part of it in the fundamental interval, or cell) are points of $C(G)$. Therefore P is an interior point of $C(G)$. It follows that $C(G)$ is an open set.

If H be an open set, a limiting point of $C(H)$ cannot belong to H. because every point of H has a neighbourhood none of the points of which belong to $C(H)$. Since every limiting point of $C(H)$ belongs to $C(H)$, the set $C(H)$ is closed.

A closed set G being essentially bounded, the complementary set $C(G)$ with respect to the unlimited interval or space in which G is contained is open in the absolute sense.

An unbounded open set will only necessarily have a closed set as its complement, provided the meaning of the term closed set is extended by admitting adjoined points at infinity as points of the closed set (see § 53). For example, consider the linear set of all points x such that $0 < x$; the complementary set with respect to the indefinite interval $(-\infty, \infty)$ is the set for which $-\infty < x \leq 0$; and this can only be regarded as a closed set if the improper point $-\infty$ is admitted as part of it.

The complement $C_G(H)$ of a closed set with respect to a perfect set G which contains H is open relatively to G. Conversely the complement with respect to the perfect set G of a set H contained in it, and open with respect to it, is a closed set.

The proof of this theorem is precisely similar to that of the last theorem.

If G_1, G_2 be closed sets, both the sets $M(G_1, G_2)$, $D(G_1, G_2)$ are closed.

Let P be a limiting point of $M(G_1, G_2)$; then P must be a limiting point of one at least of the sets G_1, G_2. For, if this were not the case, a neighbourhood of P could be determined so as to contain no points of G_1 or of G_2, other than P, and therefore none of $M(G_1, G_2)$. The point P

---

* *Intégrales de Lebesgue*, p. 106.

consequently belongs to one at least of the closed sets, and therefore to $M(G_1, G_2)$, which is therefore closed. A limiting point P, of $D(G_1, G_2)$, is clearly a limiting point both of G_1 and of G_2; therefore P belongs to both these sets, and thus to $D(G_1, G_2)$, which is consequently closed.

The theorem can be readily extended to the case of any finite number of closed sets. Thus:

If $G_1, G_2, \ldots G_n$ *be closed sets, both the sets*

$$M(G_1, G_2, \ldots G_n),\quad D(G_1, G_2, \ldots G_n)$$

are closed.

If the number of closed sets is indefinitely great, so that they form a sequence $G_1, G_2, \ldots G_n, \ldots$ of such sets, the set $M(G_1, G_2, \ldots)$ is not necessarily closed. For if P_1 belong to G_1, P_2 to G_2, P_3 to G_3, ..., the point P, the limiting point of the sequence $P_1, P_2, \ldots$, does not necessarily belong to any of the sets, and thus does not necessarily belong to $M(G_1, G_2, \ldots)$. On the other hand, the set $D(G_1, G_2, \ldots)$, if it exists, is necessarily closed; for every limiting point of it is also a limiting point of G_m, for each value of m, and is therefore a point of G_m for each value of m.

If $O_1, O_2, O_3, \ldots$ *be a finite number, or a sequence, of open sets,* $M(O_1, O_2, O_3, \ldots)$ *is also an open set.*

The sets $O_1, O_2, \ldots$ all being open sets, the complementary sets $C(O_1), C(O_2), \ldots$ are all closed sets: and consequently $D\{C(O_1), C(O_2), \ldots\}$ is a closed set, whether the number of the given sets is finite or not. But the set $D\{C(O_1), C(O_2), \ldots\}$ is clearly the complement of $M(O_1, O_2, \ldots)$, which is consequently an open set.

In case the sets $O_1, O_2, \ldots$ are all contained in a finite closed interval or cell, the complementary sets $C(O_1), C(O_2), \ldots$ are taken relatively to this interval or cell. If this is not the case, the sets $C(O_1), C(O_2), \ldots$ will only be closed, in an extended sense of the term, when limiting points at infinity are admitted. This being taken into account, the proof applies to this case.

If $O_1, O_2, \ldots O_m$ *be a finite number of open sets, the set* $D(O_1, O_2, \ldots O_m)$, *if it exists, is also an open set.*

For the set $D(O_1, O_2, \ldots O_m)$ is the complement of the set

$$M\{C(O_1), C(O_2), \ldots C(O_m)\},$$

which is a closed set.

The theorem does not hold for the case of an infinite number of open sets.

It has been seen that the properties* of a set of being perfect, closed, open, dense, or non-dense, are invariant for a wide class of transformations of which examples have been given in § 53; special account being taken of the cases in which boundaries at infinity must be adjoined to the space in which a set exists. These properties of a set are accordingly said to be descriptive properties, as distinct from the non-invariant metric properties which will be considered in Chapter III.

---

* See a paper by E. H. Neville, *Acta Math.* vol. XLIII (1920), p. 63.

PROPERTY OF THE SUCCESSIVE DERIVATIVES OF A SET

57. The following fundamental property of the successive derivatives of a set of points, in any number of dimensions, will now be established.

All the derivatives G', G'', G''', ... $G^{(n)}$, ... *of a given set are closed sets. and each of these derivatives, after the first, consists of points belonging to the preceding one, and therefore to* G'.

This theorem, usually stated for a bounded set, holds also for an unbounded set, provided the extended meaning be given to the terms, limiting point, closed set, which has been formulated in § 53 and § 55

If a point P of $G^{(n)}$, where $n \geq 2$, existed, which did not belong to G'. then a neighbourhood of P could be determined, so as to contain only a finite set of points of G, or no such points, and this neighbourhood would therefore contain no points of G', and consequently none of G'', G''', ... $G^{(r)}$; which would be contrary to the hypothesis that P belongs to $G^{(n)}$. Therefore every point of $G^{(n)}$ ($n \geq 2$) belongs to G'. By considering the case $n = 2$, we see that G' is a closed set.

If we take $G^{(n-2)}$ to be the original set, it follows from the above that every point of $G^{(n)}$, the second derivative, belongs to $G^{(n-1)}$, the first derivative. We have thus shewn that

$$G^{(n)} = D(G', G'', \ldots G^{(n)}).$$

The derivative G' *of a set* G *which is dense in itself is perfect.*

For G' is closed, and every point of G belongs to G': thus G' contains no point which is not a limiting point of G'. Therefore G', being both closed and dense in itself, is perfect.

ENUMERABLE AGGREGATES

58. *An aggregate which contains an indefinitely great number of elements is said to be enumerable*, or countable* (abzälbar, dénombrable), *when the aggregate is such that a* (1, 1) *correspondence can be established between the elements and the set of integral numbers* 1, 2, 3, ...

An aggregate of objects is therefore enumerable if the objects can be arranged in a series which has a first term and in which any assigned object belonging to the aggregate has a definite place assigned by a definite ordinal number n. Thus the elements of an enumerable aggregate can be represented by a sequence of symbols

$$u_1, u_2, \ldots u_n, \ldots.$$

---

* Cantor, *Crelle's Journal*, vol. LXXVII (1874), p. 258.

It follows from this definition that the elements of two enumerable aggregates are such that a (1, 1) correspondence can be established between them.

If a new aggregate be defined by selecting, in accordance with a rule or finite set of rules, elements from those which belong to an enumerable aggregate, an indefinitely great number of such elements being taken, then the new aggregate is also enumerable. For such an aggregate selected from $u_1, u_2, \dots u_n, \dots$ is $u_r, u_s, u_t, \dots$ $(r < s < t \dots)$, which satisfies the conditions of having a first term, and of having each element of the aggregate in a definite place in the series. It thus appears that a (1, 1) correspondence can be established between an enumerable aggregate and one which is a part of that aggregate, provided this part be not finite. This is the characteristic property which distinguishes an aggregate containing an indefinitely great number of elements from one containing only a finite number of elements. For example, a (1, 1) correspondence exists between all the integral numbers and all the odd numbers, or between all the integral numbers and all the prime numbers.

If a finite number of enumerable aggregates be given, or even if the number of such aggregates be indefinitely great, but enumerable, then the new aggregate formed by combining these aggregates into a single one is itself enumerable.

We may denote such a composite aggregate by the letters

$$\begin{array}{l}
u_{11}, \quad u_{12}, \quad u_{13}, \dots u_{1n}, \dots \\
u_{21}, \quad u_{22}, \quad u_{23}, \dots u_{2n}, \dots \\
u_{31}, \quad u_{32}, \dots\dots\dots\dots\dots\dots \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots
\end{array}$$

and we shall shew that the double sequence so formed represents an enumerable aggregate. To see this, it is sufficient to write the sequence in the form

$$\begin{array}{llll}
u_{11} \\
u_{12}, & u_{21} \\
u_{13}, & u_{22}, & u_{31} \\
\dots\dots\dots\dots \\
u_{1,n-1}, & u_{2,n-2}, & u_{3,n-3}, & \dots u_{n-1,1} \\
\dots\dots\dots\dots\dots\dots
\end{array}$$

where the sum of the indices is the same for all the terms which are written in one horizontal line. It is now clear that each number u_{pq} has a definite place in a sequence in which u_{11} has the first place; the double sequence is therefore enumerable.

An important particular case of the above result is the following theorem:

The aggregate of all the rational numbers is enumerable.

A rational number p/q may be denoted by $u_{p,\,q}$: therefore the aggregate is enumerable. It makes no difference that any particular number p/q occurs an indefinite number of times as $u_{rp,\,rq}$; since, if all such terms except those for which $r = 1$, and p/q is in its lowest terms, be removed, the aggregate left is still enumerable.

For example, the aggregate of rational numbers in the open interval $(0, 1)$ may be arranged in the order $\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \ldots$; in which each fraction that occurs is in its lowest terms.

A method of placing the rational fractions between 0 and 1 into correspondence with the integers 1, 2, 3, ... has been given by Faber*. By this method the integer corresponding to a given fraction can be calculated by means of formulae. Faber shews that there exists a unique representation of a given proper fraction p/q, in the finite form

$$\frac{a_1}{2!} + \frac{a_2}{3!} + \frac{a_3}{4!} + \ldots + \frac{a_n}{(n+1)!},$$

where a_r is an integer (or zero) $< r + 1$; whereas every integer is uniquely representable in the form

$$b_1 + b_2.2 + b_3.2.3 + \ldots + b_n.2.3.4 \ldots n.$$

where b_r is an integer (or zero) $< r + 1$.

To the proper fraction p/q there corresponds the unique integer

$$a_1 + a_2.2! + a_3.3! + \ldots + a_n.n!$$

59. A more general theorem has also been established by Cantor†. An algebraical number is one which is a root of an algebraical equation in which the coefficients are all rational numbers, so that the coefficients may without loss of generality be taken to be integers. Cantor's theorem is, that *all the algebraical numbers form an enumerable aggregate.*

To prove this theorem, let

$$p_0x^n + p_1x^{n-1} + \ldots + p_n = 0$$

be an equation in which $p_0, p_1, \ldots p_n$ are all positive or negative integers (including zero): and let

$$| p_0 | + | p_1 | + \ldots + | p_n | + n = N;$$

then N is a positive integer which may be called the rank of the equation. It is clear that there are only a finite number of equations of any given rank, these equations having only a finite number of roots. If then we let $N = 3, 4, 5, \ldots$ successively, all the algebraical numbers can be arranged in a simple sequence; and thus they form an enumerable aggregate. The aggregate which is formed of all the real algebraical numbers is consequently also itself enumerable.

---

* *Math. Annalen*, vol. LX (1905), p. 196.

† *Crelle's Journal*, vol. LXXVII (1874), p. 258.

A number which is not an algebraical number is said to be transcendental. The existence of transcendental numbers was first established* by Liouville, who shewed how examples of such numbers could be formed. No general criterion is known by which it can be decided whether a number, defined by a given analytical procedure, is algebraical or transcendental. The first case in which such a number, well known in Analysis, was shewn to be transcendental was that of the number e, the base of the natural system of logarithms; and the first proof that e is transcendental was given by Hermite. The next case in which a number, defined analytically, was shewn to be transcendental was that of the number π. The first demonstration of this important fact is due to Lindemann†, who proved the more general theorem that, if $e^x = y$, the two numbers x, y cannot both be algebraical, except in the case $x = 0$, $y = 1$. It follows that the natural logarithms of all algebraical numbers are transcendental, as also are all numbers of which the natural logarithms are algebraical.

60. The following fundamental theorem will now be established‡:

The aggregate which consists of the continuum of numbers in a given interval is not enumerable.

Suppose that $\omega_1, \omega_2, \omega_3, \ldots$ denote the numbers in an enumerable aggregate; it will then be shewn that, between any two real numbers α, β whose difference is as small as we please, a number occurs which does not belong to the enumerable aggregate. It will then follow that, in the given interval (α, β), there are an unlimited number of points which do not belong to the enumerable aggregate, and thus that the latter cannot contain all the points of the continuum. If the enumerable set of points is not everywhere dense in (α, β), then smaller sub-intervals inside (α, β) can be taken which contain no points of the aggregate; and thus we have only to consider the case in which the given aggregate is everywhere dense in (α, β). Let ω_{κ_1} be the first of the points $\omega_1, \omega_2, \ldots$ which lies within (α, β), and ω_{κ_2} be the next of these points which lies within (α, β) so that $\kappa_1 < \kappa_2$. Let α' be the smaller, and β' the greater of the numbers $\omega_{\kappa_1}, \omega_{\kappa_2}$, then $\alpha < \alpha' < \beta' < \beta$, and $\kappa_1 < \kappa_2$; and if $\mu < \kappa_2$, then ω_μ does not lie within the interval (α', β'). Considering this latter interval, let $\omega_{\kappa_3}, \omega_{\kappa_4}$ be the first two of the numbers of the enumerable aggregate which lie within (α', β'), and let α'' be the smaller and β'' the greater of these, then $\alpha' < \alpha'' < \beta'' < \beta'$, and $\kappa_2 < \kappa_3 < \kappa_4$. Proceeding in this manner, we obtain a whole series of sub-intervals each one of which is entirely within the preceding one; thus $(\alpha^{(\nu)}, \beta^{(\nu)})$ lies within $(\alpha^{(\nu-1)}, \beta^{(\nu-1)})$; and if $\mu < \kappa_{2\nu}$, then ω_μ does not lie within $(\alpha^{(\nu)}, \beta^{(\nu)})$; also

$$\kappa_1 < \kappa_2 < \kappa_3 < \ldots < \kappa_{2\nu-2} < \kappa_{2\nu-1} < \kappa_{2\nu},$$

---

* *Liouville's Journal*, vol. XVI (1851), p. 133.

† See *Math. Annalen*, vol. XX (1882), p. 213.

‡ Cantor, *Crelle's Journal*, vol. LXXVII (1874), p. 260.

and $2\nu \leqq \kappa_{2\nu}$; and thus ω_ν lies outside $(\alpha^{(\nu)}, \beta^{(\nu)})$. Since the numbers $\alpha', \alpha'', \alpha''', \ldots$ are in ascending order, and all lie within (α, β), they have a limit A; similarly $\beta', \beta'', \beta''', \ldots$ have a limit B; and $\alpha^{(\nu)} < A \leqq B < \beta^{(\nu)}$. If $A < B$, then, since all the numbers ω_ν are outside the interval (A, B), the given aggregate is not everywhere dense in (α, β); which is contrary to hypothesis. Hence we have $A \equiv B$; and the number A, or B, is a number which does not occur in the aggregate $\omega_1, \omega_2, \ldots$. thus the assumption that all the real numbers in a given interval can be effectively arranged in a simple sequence has been shewn to lead to a contradiction.

It will be observed that the point of the foregoing proof consists in the fact that an everywhere dense enumerable aggregate necessarily has limiting points which do not belong to the aggregate

A second proof[*], also due to Cantor, that the continuum is not enumerable is the following:—Without loss of generality, the interval may be taken to be (0, 1). Suppose it to be possible to state a set of rules by which all the numbers within this interval are arranged in a sequence, so that there is a first, a second, a third, and so on; and so that every number occurs somewhere in the arrangement. Since certain rational numbers are capable of double representation, viz. by means of a decimal in which, from and after some fixed place, all the digits are zero, and also by a decimal in which, from and after some fixed place, all the digits are 9, we shall suppose this last mode of representation excluded, so that each number is represented uniquely. Subject to this convention, let the numbers, in order, be exhibited as decimals

$$\begin{array}{l} \cdot p_{11}\, p_{12}\, p_{13} \cdots\cdots \\ \cdot p_{21}\, p_{22}\, p_{23} \cdots\cdots \\ \cdot p_{31}\, p_{32}\, p_{33} \cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \end{array}$$

where each p stands for one of the digits 0, 1, 2, ... 9.

It is assumed that we are in possession of a set of rules by means of which the mth digit of the nth number can be determined, for each pair of values of m and n, by means of a finite number of applications of the given set of rules. Let it be assumed that, if possible, all the real numbers in the open interval (0, 1) occur in the above sequence. If now a number in the open interval can be determined which does not occur in the above sequence, a contradiction will have been shewn to be involved in the assumption that all the numbers in the interval occur in the above sequence. Now such a number can be defined, for example, by the following rules. Let the nth digit of the number be $\bar{p}_{nn} \equiv p_{nn} + 1$, unless $p_{nn} \equiv 8$ or 9; if $p_{nn} \equiv 8$ or 9, let $\bar{p}_{nn} \equiv 0$. The number $\cdot\bar{p}_{11}\bar{p}_{22}\bar{p}_{33} \cdots \bar{p}_{nn} \cdots$

---

[*] *Jahresbericht der deutschen math. Vereinig.* vol. I (1892), p. 75.

differs in respect of at least one digit from every number in the above sequence, and therefore cannot occur in the sequence. The contradiction in the original assumption is thus established.

It will be observed that, in general, when any sequence of numbers in the interval $a < x < b$ has been defined, the existence of the sequence provides the means of defining other numbers in the interval that do not occur in the sequence.

THE POWER, OR CARDINAL NUMBER, OF AN AGGREGATE

61. A notion of fundamental importance in the theory of aggregates is that of the *power*, or *cardinal number*, of an aggregate. This notion will be considered more generally and fully in Chapter IV, where it will be shewn that the power of an aggregate is the generalization of the notion contained in the cardinal number of a finite aggregate. At present, an account of the notion of the power of an aggregate will be given, so far as it is necessary for the application to the case of sets of points.

Two aggregates of objects are said to have the same power, or cardinal number, when a (1, 1) *correspondence can be established between them, so that each element of either of the aggregates corresponds to one single element of the other.*

Finite aggregates have the same power when they consist of the same number of elements, *i.e.* when they have the same cardinal number. Of aggregates which are not finite we consider first enumerable aggregates. Every enumerable aggregate has the power of the aggregate of integral numbers; and this we may denote by a. It has been shewn above that, if from an aggregate of power a any elements be removed, then the remaining aggregate, provided it contains a non-finite number of elements, has still the same power a. It has further been shewn that the composite aggregate formed of a finite, or enumerable, set of enumerable aggregates has the same power a. It follows, as an interesting case, that the set of all those points of a p-dimensional space whose coordinates are rational numbers has the power a of the set of integral numbers, or of the rational numbers in a given linear interval.

It is easily shewn that the power of the set of all the points in an interval (a, b) is the same as that in any other finite interval, say (0, 1); for $\frac{x-a}{b-a} = \xi$ establishes a (1, 1) correspondence between the points x of (a, b) and the points ξ of (0, 1). Again, the relation $\frac{x}{\sqrt{x^2+h^2}} = \xi$ establishes a (1, 1) correspondence between all real numbers, and those within the interval $(-1, 1)$; and thus the power of all real numbers is the same as that of all those in a finite interval. This power is called the cardinal number of the continuum, and may be denoted by c.

As regards unenumerable aggregates in general, it can be shewn that the power of such an aggregate is unaltered by removing from the aggregate any elements which form an enumerable aggregate. Let A denote the given aggregate, and a the enumerable aggregate which is removed; and let B denote the remaining aggregate, which cannot be enumerable, for otherwise (a, B), or A, would be so also. From B, suppose an enumerable aggregate a' to be removed, leaving the aggregate C, thus $A = (a, a', C)$, $B = (a', C)$. Now (a, a') and a', being both enumerable, have the same power; and a $(1, 1)$ correspondence therefore exists between their elements: and since A and B have the aggregate C in common, it therefore follows that A and B have the same power As an example of this theorem, we see that the set of irrational points in a given interval has the power c of the set of all numbers in the interval. Again the set of transcendental numbers in a given interval has the power c of the continuum. whereas the set of algebraical numbers in the same interval has the power a. In this proof, the assumption has been made that the unenumerable aggregate B necessarily contains an enumerable part a'.

The known infinite sets of points defined in accordance with the methods usual in the theory of sets of points, in a line or in a continuum of any number of dimensions, have either the power a or the power c; but it has not yet been established that the assumption of the existence of an infinite set of points which has neither the power a nor the power c leads to a contradiction. Other aggregates have been contemplated which have a power higher than c; these will be referred to later. in dealing with the theory of functions.

62. *The p-dimensional continuum has the power c of the one-dimensional continuum**.

To prove this theorem, we use the fact that any irrational number in the interval $(0, 1)$ can be exhibited uniquely as an infinite continued fraction

$$x = \frac{1}{a_1+}\,\frac{1}{a_2+}\cdots\frac{1}{a_p+}\cdots,$$

where $a_1, a_2, \dots a_p, \dots$ are determinate integers for any given irrational number x in the interval $(0, 1)$. Let

$$x_1 = \frac{1}{a_1+}\,\frac{1}{a_{p+1}+}\,\frac{1}{a_{2p+1}+}\cdots,$$

$$x_2 = \frac{1}{a_2+}\,\frac{1}{a_{p+2}+}\,\frac{1}{a_{2p+2}+}\cdots,$$

....

$$x_p = \frac{1}{a_p+}\,\frac{1}{a_{2p}+}\,\frac{1}{a_{3p}+}\cdots,$$

---

* Cantor, *Crelle's Journal*, vol. LXXXIV (1878), p. 242.

thus, corresponding to any value of x, a set of irrational numbers $x_1, x_2, \ldots x_p$ is uniquely determined, and conversely, corresponding to any set of irrational numbers $x_1, x_2, \ldots x_p$, an irrational number x is uniquely determined.

It has thus been shewn that the irrational points of the linear continuum (0, 1) correspond uniquely to those points of the p-dimensional continuum in which each coordinate is in the interval (0, 1), and is irrational. It has been shewn in § 61 that the set of irrational values of x_1, in the interval (0, 1), has the same power as the set of all the numbers in this interval. Since this holds also for $x_2, x_3, \ldots x_p$, it follows that a (1, 1) correspondence exists between that set of points in the n-dimensional continuum, for which $x_1, x_2, \ldots x_p$ all have irrational values, and the set in which $x_1, x_2, \ldots x_p$ have all values rational or irrational; thus these sets have the same power. Hence the set of all points of the p-dimensional continuum, in which each coordinate is in the interval (0, 1), has the same power as the set of all points in the linear interval (0, 1). It has thus been shewn that the p-dimensional continuum has the same power c as that of one dimension.

THE ARITHMETIC CONTINUUM

63. The arithmetic continuum being regarded as obtained by adjoining to the set of rational numbers the set of all their limiting points, the question arises how far it is legitimate to consider the complete set so obtained as constituting a single object, determined by means of the elements of which it is composed. A finite set of numbers, or points, constitutes a single object determined by means of its parts, in the sense that those parts can be exhaustively exhibited by means of a finite number of specifications representable by a finite number of symbols. An enumerable set of numbers, or of points, in particular the set of rational numbers, is not determinate in the sense that the elements of the set can be exhaustively exhibited; but it is determinate in the sense that a rule of tabulation can be given, such that each particular number of the set occupies a determinate place in the table; and each particular number can be represented by means of a finite number of symbols. Such a set may be regarded as an aggregate, or single object, in the same sense in which the natural numbers 1, 2, 3, ... may be regarded as forming an aggregate (see § 4). When we come, however, to the case of the continuum, or aggregate of all real numbers, the fact that this aggregate is unenumerable introduces a new element into the question of the legitimacy of considering the set of these numbers as forming a determinate whole, or as constituting a single object of thought. The set of real numbers cannot be tabulated in such a manner that no number fails to occur at some definite place in the table. In fact it has been shewn, in § 60, that the assumption of the existence of such a table leads to contradiction. It thus appears that no set of rules or specifications can be given which suffice to determine successively all the numbers of the set; and no finite set of symbols can exhaustively exhibit the numbers. The only sense in which the numbers of the set are determinate is that each such number is the limit of a convergent sequence of numbers, taken from the unending table formed by the rational numbers. It may fairly be doubted whether such a negative specification of elements amounts to a valid synthetical definition of a determinate aggregate; this point will however be further discussed in Chapter IV, in connection with the general theory of aggregates. It will there be shewn that the arithmetic continuum has an order-type possessing definite characteristics which, in their totality. uniquely characterize it. This expresses the only kind of unity which can appertain to the continuum, considered as a synthetic arithmetic construction. From the point of view of Arithmetical Analysis the existence of the aggregate of real numbers, as a single definite object, possessing assigned properties, may be regarded as a fundamental postulate, the validity of such postulation being subject to the law of contradiction If it be held that we possess an independent knowledge of the existence of the geometrical continuum, derived by a process of idealization from our intuition of space, we may regard the function of the set of real numbers to consist. not in a synthetical formation of the concept of the continuum, but inversely in an analysis of the contents of the continuum It is difficult to see how precision can be introduced into the intuitional notion of the spatial continuum apart from some theory relating either to points or to infinitesimals; and the language employed in such a theory must be of a symbolical character amounting to the use of some kind of arithmetical notation Regarding the geometrical continuum in this way as a single object of which we have a direct knowledge obtained from our intuitions of space and time, the reduction to a precise abstract form may be regarded as being made upon the assumption that the system of rational numbers, with their limits adjoined, is adequate to the analytical description of the continuum, in the sense that each point in the continuum is represented uniquely by a single real number, and that there is no point in the continuum which is not so represented. This amounts to a definition, in a certain sense, of the contents of the geometrical continuum. Such definition is not the only possible definition, but it is a legitimate one, provided it suffices for the purposes we have in view in Analysis and Geometry, and provided it does not conflict with the concept of Continuity as derived from intuition. The *generic* distinction between a continuous geometrical object, and a point, or set of points, situated in that object, is not capable of direct arithmetic representation. This does not, however, impair the efficiency of Arithmetical Analysis in dealing with geometrical objects. In Cartesian geometry, for example, Analysis is really concerned only with the points that can be determined in the geometrical objects with which it deals. This does not mean that a continuous geometrical object is analysed into points which are of necessity to be regarded as its "parts," but it does mean that, for the particular purposes of Analytical Geometry, an adequate treatment of it is to regard it as a set of points

Taking as data the set of rational numbers, an irrational number of the continuum is defined in some manner which involves the use of words and symbols. Such a symbol may have a unique meaning, or it may be of the nature of a variable having as its field some enumerable set of rational, or of integral, numbers already defined, for example the symbol n may be taken to denote any integer of the sequence of natural numbers By means of a given stock of such words and symbols it is possible to define only an enumerable set of elements of the continuum, each word, or symbol, being employed a finite number of times only. On this ground it has been maintained* by König that the elements of the arithmetic continuum fall into two classes; the first class consisting of numbers capable of finite definition, and those of the second class being inherently incapable of finite definition. If this view were justifiable, grave difficulties would arise in the whole theory of the arithmetic continuum, which is the basis upon which Arithmetic Analysis rests. For an object that is inherently incapable of being defined finitely may be held not to be definable at all, and such an object is regarded by many, if not most, mathematicians as not being an existent object for mathematical thought. The fallacy involved in the introduction of this distinction appears clearly when the method and implications of Cantor's second proof, given in § 60, that the continuum is unenumerable, are fully scrutinized. It was there shewn that if, by means of a given stock of words and symbols, a set of numbers forming an enumerable aggregate is defined, the existence of such aggregate enables us to introduce a new word or symbol, or to give a new meaning to an existing symbol, which will represent the enumerable aggregate itself, and can then be employed for the purpose of defining new elements of the continuum which do not occur in the aggregate. Such a new element has a finite definition in the same sense as that in which the elements of the enumerable aggregate in question have finite definitions. It must be remembered that an enumerable aggregate may itself be denoted by a single word or symbol that can be created or assigned *ad hoc*, whenever the enumerable aggregate has been defined.

The essence of the proof in § 60 is then that there exists, and can exist, at any time, no stock of words and symbols which cannot be increased for the purpose of defining new elements of the continuum. The theorem that the continuum is unenumerable is equivalent to the fact that the assumption of the existence of a *final and complete* stock of words and symbols, by means of which alone elements of the continuum can be defined by finite definitions, leads to a contradiction. Those elements of the continuum which are at the present time, or will be at a future time, capable of definition by means of the words and symbols in existence, may form an enumerable set; but this does not prove that there exists any element of the continuum that is inherently incapable of finite definition. In fact it can be proved that those elements of the continuum that remain when the hypothetical elements incapable of finite definition are removed, form an aggregate which has all the properties of the continuum, and can therefore be identified with it.

---

* See Konig, *Math. Annalen*, vol. LXI (1905), p. 156, also vol. LXIII (1907), p. 217. See also Richard, *Acta Math.* vol. XXX (1906), p. 295, and Hobson, *Proc. Lond. Math. Soc.* (2), vol. IV (1907), p. 210, and Whitehead and Russell, *Principia Mathematica*, vol. I (1910), p. 64.

Let us consider the continuum of real numbers in the interval (0, 1), and let G denote the set of those numbers in it that are capable of finite definition. In the first place, the set G is dense in itself, for any element of it is the limit of a sequence of rational numbers, all of which belong to G. Next, the set G is closed; for if every number of a convergent sequence $\{x_n\}$ belong to G, it can be proved that the limit of the sequence belongs to G. Denoting the set $\{x_n\}$ by E, the number which is the limit of the sequence can be finitely defined as follows: Let the numbers of E be represented by non-terminating decimals so that in none of them are all its digits, from and after a fixed one, equal to 9. Let the m^{th} digit of a number x be defined as that digit which is identical with the m^{th} digit of an infinite number of the elements of E. The number x so defined is the limit of the sequence of numbers all of which belong to G, and x, being thus finitely defined, itself belongs to G.

The set G is thus dense in itself and closed, and is therefore perfect: it is clearly everywhere dense in the interval (0, 1). It has therefore all the properties of the continuum, and therefore the hypothetical non-definable elements can be disregarded.

It will be observed that the elements of the continuum cannot be exhaustively represented by any finite set of symbols, each used a finite number of times. Thus there can exist no *systematic* notation which suffices to represent all the numbers of the continuum.

TRANSFINITE ORDINAL NUMBERS

64. The theory of transfinite ordinal numbers had its origin* in the investigation of the theory of sets of points. The general abstract theory of such numbers, or order-types, will be deferred until Chapter IV; it is necessary however to introduce here the conceptions connected with the formation of these numbers, with a view to utilizing them in the theory of sets of points.

---

* An account of Cantor's earliest presentation of this subject will be found in *Math. Annalen*, vol. XXI (1883), p. 545.

Let $P_1, P_2, \dots P_n, \dots$ denote a sequence of points in a given interval AB, representing a sequence $a_1, a_2, a_3, \dots$ of increasing numbers, so that

$$a_1 < a_2 < a_3 < \dots < a_n < \dots.$$

This sequence of points has a limiting point which is not one of the points of the sequence, and is on the right of all those points; this limiting point we may denote by P_ω. The symbol ω may be regarded as denoting a new ordinal number which comes after all the ordinal numbers $1, 2, 3, \dots n, \dots$; it is called the *first transfinite ordinal number*. The number ω is not contained in the sequence of finite ordinal numbers, but comes after all of them; and we shall see that it may be taken as the first of a new sequence of ordinal numbers, all of which must be regarded as ordinally greater than the finite ordinal numbers.

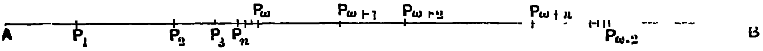

FIG. 1

Suppose that beyond the point P_ω there are other points which we wish to regard as belonging to the same set as the points $P_1, P_2, \dots P_n, \dots P_\omega$; then these points will be denoted by $P_{\omega+1}, P_{\omega+2}, \dots P_{\omega+n}, \dots$; and if these points are finite in number, there will be one of them $P_{\omega+m}$ which is the last on the right. The indices of all the points of the set will then be

$$1, 2, 3, \dots n, \dots \omega, \omega+1, \omega+2, \dots \omega+m;$$

and the numbers $\omega, \omega+1, \dots \omega+m$ are regarded as a set of transfinite ordinal numbers, which commences with the first transfinite ordinal number ω, and contains the m succeeding transfinite ordinal numbers. It may however happen that the set of points $P_\omega, P_{\omega+1}, P_{\omega+2}, \dots$ has no last point. In that case, assuming that the points are all contained in a finite interval, the set has a limiting point which is not contained in the set itself; and this limiting point we denote by $P_{\omega+\omega}$ or $P_{\omega.2}$, where $\omega.2$ is an ordinal number which is not contained in the set $\omega, \omega+1, \omega+2, \dots$, but comes after the numbers of that set.

If we wish to include further points which are on the right of $P_{\omega.2}$, we must introduce numbers denoted by $\omega.2+1, \omega.2+2, \dots$; and, in case these points form an infinite set in a finite interval, they will have a limiting point which will be denoted by $P_{\omega.2+\omega}$ or $P_{\omega.3}$. We have now the ordinal numbers

$$1, 2, 3, \dots \omega, \omega+1, \omega+2, \dots \omega.2, \omega.2+1, \omega.2+2, \dots \omega.3.$$

If we proceed further in this manner, it is clear that we may require numbers $\omega.n, \omega.n+1, \omega.n+2, \ldots \omega.\overline{n+1}, \ldots$, where n denotes any finite number.

Further, it may happen that the set of points $P_\omega, P_{\omega.2}, P_{\omega.3}, \ldots P_{\omega.n}, \ldots$ is itself infinite, and has a limiting point on the right of all these points. This point we denote by P_{ω^2}; and the number ω^2 we consider to be a new ordinal number which succeeds all the numbers $\omega.n+m$, where n and m have all possible finite values.

Points on the right of P_{ω^2} may be denoted by means of the indices $\omega^2+1, \omega^2+2, \omega^2+3, \ldots$; and if these points are infinite in number, they may have a limiting point $P_{\omega^2+\omega}$.

Points on the right of $P_{\omega^2+\omega}$ may be denoted by the indices $\omega^2+\omega+1$, $\omega^2+\omega+2, \ldots$; if these have a limiting point, it will be denoted by the index $\omega^2+\omega.2$. Proceeding in this manner, we may have points of which the indices are $\omega^2+\omega.3, \omega^2+\omega.4, \ldots$. If there is an infinite set of such points, and the set has a limiting point, on the right of the set, this limiting point will have $\omega^2+\omega^2$, or $\omega^2.2$, for its index.

If we proceed still further, we see as before that we may have to contemplate numbers of the form $\omega^2.p+\omega.q+r$, where p, q, r are finite: afterwards $\omega^3, \omega^3+1, \ldots, \omega^3.p+\omega^2.q+\omega.r+s$, &c. The general type of ordinal numbers which can be obtained in this manner is represented by $\omega^n.p_n+\omega^{n-1}.p_{n-1}+\ldots+\omega.p_1+p_0$, and it is clear that, for the representation of points of a given set, such numbers may be required as indices.

It may happen that the set of points whose indices are $\omega, \omega^2, \omega^3, \ldots$ is not finite; then the limiting point of such set will be denoted by the index ω^ω. Starting afresh with this number, we may form numbers such as

$$\omega^{\omega^n.p_n+\omega^{n-1}.p_{n-1} \qquad p_0}.$$

If the points whose indices are $\omega^\omega, \omega^{\omega^2}, \omega^{\omega^3}, \ldots$ do not form a finite set, their limiting point will be denoted by ω^{ω^ω}.

In a similar manner we may denote by ϵ_1 the number which comes after the sequence $\omega, \omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}} \ldots$; and starting from ϵ_1, we may similarly proceed to form further numbers in endless succession.

65. All the ordinal numbers which can be formed in the manner above described are formed by means of the application of Cantor's two principles of generation (Erzeugungsprinzipien).

(1) *After any number another immediately succeeding it is formed by the addition of unity.*

(2) *After any endless sequence of numbers, a new number is formed which succeeds all the numbers in the sequence, and has no number immediately preceding it.*

All transfinite ordinal numbers which can be formed by means of these two principles of generation are said to be ordinal numbers of the *second class*. The finite ordinal numbers are said to be of the *first class*, they are formed successively, starting with the number 1, by means of the first principle of generation alone.

The numbers of the second class are of two essentially distinct species: (1), *non-limiting numbers*, those numbers which have each a number immediately preceding them, and from which they are formed by the addition of unity; for example $\omega + n$, $\omega^2 . p + \omega . q + 1$, $\omega^\omega + \omega + 1$ and (2), *limiting numbers*, those which have no number immediately preceding them, from which they are formed by the addition of unity, for example ω, $\omega^2 + \omega$, $\omega^\omega + \omega^2$ are limiting numbers.

Any particular number of the second class can be denoted by a finite number of symbols, but there is no upper limit to the number of symbols required to denote such numbers.

Cantor has further postulated the existence of a number Ω which comes after all the numbers of the second class, and is the first number of a new set which is called the third class. The number Ω cannot be obtained as the number which succeeds a simple sequence, by means of the second principle of generation; for every number which can be so obtained is itself a number of the second class. This number Ω can be obtained only by means of a third principle of generation, which postulates the existence of a new number coming after all the numbers of the complex formed by the application of the first and second principles of generation. The validity of the postulation of the existence of the number Ω, and of the higher numbers of the third class, will be discussed in Chapter IV.

66. A fundamental property of the numbers of the second class may be expressed as follows:

Let $P_1, P_2, P_3, \ldots P_n, \ldots P_\omega, P_{\omega+1}, \ldots$ *be an infinite set of points such that either* (1), *there is a last point* P_β, *where* β *is some number of the second class, or* (2), *there is no last point, but every index occurs which is less than some limiting number* γ *of the second class, whereas the index* γ *itself does not occur; the set of points is then enumerable.*

The sets

$$\begin{array}{lll}
P_1, & P_2, & \ldots P_n, \ldots \\
P_\omega, & P_{\omega+1}, & \ldots P_{\omega+n}, \ldots \\
P_{\omega.2}, & P_{\omega.2+1}, & \ldots P_{\omega.2+n}, \ldots \\
P_{\omega.3}, & P_{\omega.3+1}, & \ldots \\
\cdots\cdots & & \\
P_{\omega.r}, & P_{\omega.r+1}, & \ldots \\
\cdots\cdots & &
\end{array}$$

where every index less than ω^2 occurs, form an enumerable aggregate of enumerable sets of points; and this has been shewn, in § 58, to be itself an enumerable set. Now consider the sets

$$\begin{array}{llll}
P_1, P_2, P_3, & \dots P_\omega, & P_{\omega+1}, \dots P_{\omega.2}, \dots P_{\omega.3}, \dots \\
P_{\omega^2}, & P_{\omega^2+1}, & \dots P_{\omega^2+\omega}, & P_{\omega^2+\omega+1}, \dots \\
P_{\omega^2.2}, & P_{\omega^2.2+1}, & \dots P_{\omega^2.2+\omega}, \dots & \\
P_{\omega^2.3}, & P_{\omega^2.3+1}, & \dots P_{\omega^2.3+\omega}, \dots & \\
\dots\dots\dots & \dots\dots\dots & \dots\dots & \\
\dots\dots\dots & \dots\dots\dots & \dots\dots &
\end{array}$$

such that in the first set there is every index less than ω^2, in the second, every index less than $\omega^2.2$, and so on. Each of these sets is enumerable, and there is an enumerable set of such sets; hence the whole set, which contains every index less than ω^3, is enumerable. In this manner it can be shewn that, if every index less than ω^n occurs, the set is enumerable. If the theorem holds for sets which contain every index less than β_1, β_2, β_3, ..., then it holds for a set which contains every index less than β, the limiting number of the sequence $\beta_1, \beta_2, \beta_3, \dots$. For the points with indices less than β_1, with indices $\geqq \beta_1$ and $< \beta_2$, with indices $\geqq \beta_2$ and $< \beta_3$, &c. form an enumerable sequence of enumerable sets, therefore, by the theorem of § 58, the whole set with indices $< \beta$ is enumerable. Since the theorem holds for $\beta_1 = \omega$, $\beta_2 = \omega^2$, $\beta_3 = \omega^3$, ... it holds for $\beta = \omega^\omega$. By continual application of this method, since any number can be reached by means of the two principles of generation, and since every number is either a limiting number, or is obtained from one by adding a finite number, we see that the general theorem holds

It will now be shewn, conversely, that *if a set of points* $P_1, P_2, \dots P_n$, $\dots P_\omega, \dots P_\beta, \dots$ *is enumerable, there must be some definite number* γ *of the first or of the second class, such that* γ *does not occur among the indices of the points, and such that every number less than* γ *does so occur.*

In case γ is a limiting number, there is no last point of the set; but if γ is not a limiting number, there is a last point, viz. the one of which the index is the number immediately preceding γ.

To prove the theorem, we observe that, since the given set of points is enumerable, it may be placed in correspondence with a set of points $Q_1, Q_2, \dots Q_n, \dots$ in which all the indices are numbers of the first class. Let us suppose that, if possible, no number γ exists; and let P_{α_1} be the point of $\{P\}$ which corresponds to the point Q_1 of $\{Q\}$. Let Q_{x_1} be the point of $\{Q\}$ of smallest index, such that the corresponding point of $\{P\}$ has an index which is $> \alpha_1$; denote this index by α_2. Then let Q_{x_2} be that point of $\{Q\}$, of smallest index, such that the corresponding point of $\{P\}$ has an index $> \alpha_2$; denote this index by α_3. Proceeding in this manner, we have a set of points $Q_1, Q_{x_1}, Q_{x_2}, \dots Q_{x_n}, \dots$ corresponding in order to a set of points $P_{\alpha_1}, P_{\alpha_2}, P_{\alpha_3}, \dots P_{\alpha_n}, \dots$ where $\alpha_1 < \alpha_2 < \alpha_3 < \dots < \alpha_n < \dots$. There exists a number α of the second class, which is the limit of the sequence $\alpha_1, \alpha_2, \dots \alpha_n, \dots$; and by hypothesis there exists a point P_α, which has α for index. Now the set $\{Q\}$ can contain no point which corresponds to P_α, because each point Q_n corresponds to a point of $\{P\}$ with an index less than α, and thus there is a contradiction in the hypothesis that α occurs amongst the indices of the points of $\{P\}$. Hence there exist numbers of the second class which do not occur as indices in the set $\{P\}$, and these numbers form a set which is a part of the aggregate of numbers of the second class. In this set there must be a lowest number γ and this number γ is the first which does not occur amongst the indices of the set $\{P\}$. That every part of the aggregate of numbers of the first and second classes has a lowest number will be shewn, in Chapter IV, to be a consequence of the structure of the ordered aggregate.

EXAMPLES

1. On a straight line AB, let us denote by $P_1, P_2, P_3, \dots$, those points at which the ratio AB/PB has the values 1, 2, 3, .. . The point P_1 coincides with A, and the point B can only be represented by P_ω. Now take any one of the segments P_rP_{r+1}, this may for convenience be represented on an enlarged scale. Denote by $Q_{r1}, Q_{r2}, Q_{r3}, \dots$, the points on P_rP_{r+1}, at which P_rP_{r+1}/QP_{r+1} takes the values 1, 2, 3, .. ; thus P_{r+1} can only be represented by $Q_{r\omega}$. Supposing this to have been done with every segment P_rP_{r+1} of AB, let us imagine all the points Q to be marked on AB, and to be numbered from left to right.

In P_1P_2, we shall have $1, 2, 3, \dots \omega$,

in P_2P_3 there will be $\omega+1, \omega+2, \dots \omega.2$,

and in P_3P_4 $\omega.2+1, \omega.2+2, .. \omega.3$;

the point B can be represented only by ω^2. If now we proceed to take each segment $Q_{rs}Q_{r,s+1}$, and to divide it in a similar manner, at points R for which $Q_{rs}Q_{r,s+1}/RQ_{r,s+1}$ has the values 1, 2, 3, ..., and then imagine all the points R obtained in every such segment $Q_{rs}Q_{r,s+1}$ to be marked on AB, and numbered as before, from left to right, it will be seen that all the numbers $\omega^2p+\omega q+r$ will be required, and that the point B can be represented by ω^3. The points $P_1, P_2, \dots P_\omega$ will have for their ordinal numbers

$$1, \omega^2, \omega^2.2, \omega^2.3, \dots \omega^3;$$

the point Q_{rs} will be numbered $\omega^2.r+\omega.s$; the finite numbers are all used up in the first sub-segment of AB. By proceeding to further subdivision, we may exhibit on AB, the ordinal numbers $\omega^n p_n + \omega^{n-1}p_{n-1} + \dots + p_0$, and the point B will then be represented by ω^{n+1}.

2. The properties of the integral numbers in relation to their prime factors may be employed to rearrange the series 1, 2, 3, ..., so that the numbers may be made to correspond with a series of ordinal numbers of the first and second classes.

First take the primes 1, 2, 3, 5, 7, 11, ...; these correspond with the numbers of the first class 1, 2, 3, ... n, ... Then take the squares of the primes, omitting unity; we thus have $2^2, 3^2, 5^2, 7^2, 11^2, \dots$, corresponding to $\omega, \omega+1, \omega+2, \dots \omega+n, \dots$.

We then take the cubes of the primes,

$2^3, 3^3, 5^3, 7^3, 11^3, \dots$, corresponding to $\omega.2, \omega.2+1, \dots \omega.2+n, \dots$, and in general, 2^{r+1}, 3^{r+1}, 5^{r+1}, ..., corresponding to $\omega.r$, $\omega.r+1$, .. $\omega.r+n$, ... We may then take the numbers ab which consist of the product of two prime factors; these, arranged in ascending order, correspond to ω^2, ω^2+1, ... ω^2+n, ... Next take the numbers a^2b^2, which consist of the squares of the last set; these correspond to $\omega^2+\omega$, $\omega^2+\omega+1$, ... We then take the successive sets of numbers of the forms a^3b^3, a^4b^4, .. ; we thus obtain the numbers which may be taken to correspond with

$$\omega^2+\omega.2,\ \omega^2+\omega.2+1,\ \ldots\ \omega^2+\omega.3,\ \ldots\ \omega^2+\omega.p+q,\ \ldots,$$

all of which are less than $\omega^2.2$. The sets of numbers of the forms

$$a^2b,\ (a^2b)^2,\ ..\ (a^2b)^n,\ ..\ a^3b,\ (a^3b)^2,\ ..\ (a^3b)^n,\ ..\ a^4b,\ (a^4b)^2,\ ..\ ,$$

may then be taken. Afterwards, we may proceed with the numbers which contain three different prime factors, and so on. It is clear that this mode of rearranging the integral numbers in their natural order, so that they correspond in the new order with ordinal numbers of the first and second classes, admits of great variety. In every case, there will be some lowest number of the second class, which is not employed in the correspondence established.

PROPERTIES OF AGGREGATES OF CLOSED SETS

67. The following theorem, due to Cantor, relating to a sequence of closed sets, each of which contains the next, will be established.

Having given a sequence of closed sets G_1, G_2, ... G_m, .. each of which contains the next, there exists a closed set G_ω each point of which belongs to G_m, for every value of m.

The sets may be in any number of dimensions.

Apply to the fundamental cell (or interval) in which G_1 is contained a system of nets D_1, D_2, ... D_n, ... There must be at least one mesh d_1, of the net D_1, that contains points which belong to all of the sets G_m; for otherwise, for some value of m, no mesh would contain a point of G_m. In case more than one mesh of D_1 has this property, we may suppose d_1 to be that one of such meshes which has lowest rank in the ordered set of all the meshes of D_1. Consider next those meshes of D_2 that are in d_1; as before there must be at least one of such meshes which contains points that are in all the sets G_m; let d_2 be that one of such meshes which has lowest rank of all the meshes of D_2 that are in d_1. Proceeding in this manner, there is determined a sequence d_1, d_2, ... d_m, ..., of meshes belonging to D_1, D_2, ... D_m, ... respectively, each of which contains the next, and each of which contains points that belong to G_m for every value of m. The sequence $\{d_m\}$ defines uniquely a point P that is in all the meshes of the sequence. This point P is a limiting point of G_m, for each value of m; and since G_m is closed, it belongs to G_m. Thus the existence of at least one point of G_ω has been established. The set G_ω is possibly finite, and is in any case closed. For, in an arbitrarily small neighbourhood of a limiting point P of the set G_ω, there exist points of G_ω, and therefore points of G_m; hence P is a limiting point of G_m. Since G_m is closed, P must be a point of G_m for each value of m, and therefore P belongs to G_ω.

The corresponding statement does not necessarily hold good for a sequence $\{G_m\}$ of sets, each of which contains the next, when the sets G_m are not closed; for the point P, in the above proof, is then not necessarily a point of the unclosed set G_m. The set G_ω therefore does not necessarily exist.

This is illustrated by the case in which G_m is the linear set of points $\frac{1}{m}, \frac{1}{m+1}, \frac{1}{m+2}, \ldots$. The point zero is such that, in any neighbourhood of it, points of G_m exist for each value of m, but that point does not belong to G_m; consequently the set G_ω does not exist.

The following generalization of Cantor's theorem has been given* by Sierpiński:

Having given a family F of closed sets, the necessary and sufficient condition that there exists one point at least which belongs to all the closed sets G, of F, is that every finite number of the sets G belonging to the family F has at least one point in common.

This may be proved by the use of a system of nets

The following is a corollary of this theorem:

If F is a family of closed sets G, such that, if any pair of the sets G be taken, one of them contains the other, there exists at least one point common to all the sets G which belong to the family F.

THE TRANSFINITE DERIVATIVES OF A SET OF POINTS

68. If G denotes a set of points in a cell (or interval) (a, b), it has been shewn in § 57 that the derivatives $G^{(1)}, G^{(2)}, \ldots G^{(n)}, \ldots$ are all closed sets, and that all the points of any one of these sets, after the first, are contained in the preceding set. If G is of the second species, $G^{(n)}$ exists for all values of n, and in this case, in accordance with the theorem of § 67, the set $D(G^{(1)}, G^{(2)}, \ldots)$ which contains points belonging to every $G^{(n)}$ exists, and is closed. This closed set may be denoted by $G^{(\omega)}$, where ω is the first transfinite number. It is defined to be the derivative† of G, of order ω. In case $G^{(\omega)}$ contains more than a finite set of points, we can proceed to form successive derivatives of $G^{(\omega)}$ in a manner similar to that in which the derivatives $G^{(1)}, G^{(2)}, \ldots$ of G were formed. These successive derivatives may be denoted by $G^{(\omega+1)}, G^{(\omega+2)}, \ldots G^{(\omega+n)}, \ldots$, and are regarded as the derivatives of G of orders $\omega+1, \omega+2, \ldots \omega+n, \ldots$. They have the same properties as the derivatives of finite order, viz. that all the points of each are points of $G^{(1)}$, and that all the points of any one of them are points of the preceding ones.

It may happen that one of the derivatives $G^{(\omega+n)}$ contains no points;

---

* *Bulletin of Acad. of Sciences of Cracow*, April—May, 1918.

† See Cantor, *Math. Annalen*, vol. XVII (1880), p. 357.

then the process of forming derivatives has come to an end, the last one being $G^{(\omega+n-1)}$. If this is not the case, a repetition of the above reasoning shews that the set $D\,(G^{(\omega+1)}, G^{(\omega+2)}, \ldots G^{(\omega+n)}, \ldots)$ contains at least one point, and is a closed set; this set is denoted by $G^{(\omega\cdot 2)}$, and is defined to be the derivative of G of order $\omega.2$. In the same manner we can proceed to form further derivatives, whose orders are numbers of the second class.

In general, if $\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n, \ldots$ denote a sequence of numbers of the second class, whose limiting number is β, the same reasoning as before shews that, if all the derivatives $G^{(\alpha_1)}, G^{(\alpha_2)}, \ldots G^{(\alpha_n)}, \ldots$ exist, then the set $D\,(G^{(\alpha_1)}, G^{(\alpha_2)}, \ldots G^{(\alpha_n)}, \ldots)$ contains at least one point, and is a closed set. This is denoted by $G^{(\beta)}$, and is defined to be the derivative of G of order β.

If we form the successive derivatives of the set G, whose orders are the numbers of the first and second classes, it may happen that there is a first number γ, of the first or second class, for which $G^{(\gamma)} \equiv 0$; but this number γ cannot be a limiting number of the second class.

It may, however, happen that no number γ, of the first or second class, exists for which $G^{(\gamma)} \equiv 0$, so that derivatives of G exist of orders corresponding to all the numbers of the first and second classes. It will be shewn in § 82 that, if $G^{(\gamma)}$ does not vanish, for some number γ, of the first or of the second class, then there necessarily exists a number β, of the first or second class, such that $G^{(\beta)} \equiv G^{(\beta+1)} \equiv G^{(\beta+2)} \equiv \ldots$. This set $G^{(\beta)}$ is a perfect set, and it is frequently denoted by $G^{(\Omega)}$, where Ω is the first transfinite number of the third class. The notation $G^{(\Omega)}$ may however be employed, independently of the acceptance of the theory of numbers of the third class.

Conversely, if $G^{(\Omega)}$ does not exist, $G^{(\gamma)}$ must first vanish for some number γ of the first or second class, which number cannot be a limiting number.

EXAMPLES

1[*]. Let G denote the enumerable set of points, each one of which is given by

$$\frac{1}{2^n} + \frac{1}{2^{n+m_1}} + \frac{1}{2^{n+m_1+m_2}} + \ldots + \frac{1}{2^{n+m_1+m_2+\ldots+m_n}},$$

where n has all positive integral values, excluding zero, and $m_1, m_2, \ldots m_n$ have all positive integral values including zero, independently of one another.

It is easily seen that in $G^{(n)}$, the points $\frac{1}{2^n}, \frac{1}{2^{n+1}}, \ldots$ all occur, and hence that $G^{(\omega)}$ exists, and consists of the single point zero.

2[*]. Let G denote the enumerable set of points, each one of which is given by

$$\frac{1}{2^{m_1}} + \frac{1}{2^{m_1+m_2}} + \ldots + \frac{1}{2^{m_1+m_2+\ldots+m_n}} + \frac{1}{2^{m_1+m_2+\ldots+m_n+p}} + \frac{1}{2^{m_1+m_2+\ldots+m_n+p+q_1}}$$
$$+ \frac{1}{2^{m_1+m_2+\ldots+m_n+p+q_1+q_2}} + \ldots + \frac{1}{2^{m_1+m_2+\ldots+m_n+p+q_1+q_2+\ldots+q_p}},$$

---

[*] These examples were given by Mittag-Leffler, *Acta Math.* vol. IV (1884), p. 58.

where $m_1, m_2, \ldots m_n, p, q_1, q_2, \ldots q_p$ have all positive integral values, including zero. In this case $G^{(\omega+n)}$ consists of the single point zero.

3*. Let G denote the enumerable set of points, each one of which is given by

$$\frac{1}{2^n}+\frac{1}{2^{n+m_1}}+\frac{1}{2^{n+m_1+m_2}}+\ldots+\frac{1}{2^{n+m_1+\ldots+m_n}}+\frac{1}{2^{n+m_1+\ldots+m_n+p}}$$

$$+\frac{1}{2^{n+m_1+\ldots+m_n+p+q_1}}+\ldots+\frac{1}{2^{n+m_1+m_2+\ldots+m_n+p+q_1+q_2+\ldots+q_p}},$$

where $n, m_1, m_2, \ldots m_n, p, q_1, \ldots q_p$, have all positive integral values. In this case $G^{(\omega 2)}$ exists, and consists of the single point zero.

SETS OF INTERVALS OR CELLS

69. The properties of a set of intervals, or of cells, which intervals, or cells, are assigned in any manner, are closely connected with the properties of sets of points, and will therefore be considered here in some detail.

Let us suppose that a finite set of non-overlapping intervals has been defined, all in the finite interval (a, b). Denoting these intervals by (a_1, b_1) (a_2, b_2) ... (a_n, b_n), let us suppose that all the intervals (a_r, b_r) of the set are such that none of their lengths $b_r - a_r$ are less than a fixed positive number ϵ. Then the number n of the intervals cannot exceed $(b-a)/\epsilon$. For the $2n$ points $a_1, b_1, a_2, b_2, \ldots a_n, b_n$ divide (a, b) into at most $2n+1$ parts, the sum of the lengths of which is $b-a$. These parts consist of the n given intervals, and in general of complementary intervals; it follows that the sum of the lengths of the n given intervals cannot exceed $b-a$; and thus that the number of the intervals cannot exceed $(b-a)/\epsilon$.

Next, suppose that any non-finite set of non-overlapping intervals is defined, all of the intervals lying in the fundamental interval (a, b). Choose a sequence $\epsilon_1, \epsilon_2, \ldots \epsilon_n, \ldots$ of positive decreasing numbers which converges to zero. The number of intervals of the given set which are of length $\geq \epsilon_n$ is finite, since it does not exceed $(b-a)/\epsilon_n$. The intervals of the given set can now be arranged in order of their lengths, taking first those that are $\geq \epsilon_1$, then those that are $< \epsilon_1$ and $\geq \epsilon_2$; and so on; there being only a finite number of intervals in each set. In case a number of intervals are of equal length their order may be that in which they occur in (a, b), from left to right. Therefore, since all the intervals of the given set can be arranged as a simply infinite aggregate, they form an enumerable aggregate.

Next, suppose that the intervals of a given non-overlapping set are in the unbounded interval $(-\infty, \infty)$, in which the position of any point is denoted by x. If we consider the correspondence $\xi = x/\sqrt{x^2+1}$, where

---

* This example was given by Mittag-Leffler, *Acta Math.* vol. IV (1884), p. 58.

the radical always has the positive sign, the unlimited interval $(-\infty, \infty)$ corresponds to the open segment $(-1, 1)$ in which the point ξ lies. The intervals of the given set correspond uniquely to intervals in $(-1, 1)$, and the set of these latter intervals is enumerable; hence the given set is so also. It has thus been shewn that:

*Every set of intervals in a bounded, or unbounded, interval which is such that no two of the intervals overlap is either finite, or forms an enumerable aggregate**.

The corresponding theorem can be shewn to hold for a set of non-overlapping cells in two or more dimensions. Two such cells may have a portion of their boundaries in common.

It will be sufficient to consider the case of plane cells; the method of proof can be extended to the case of cells of any number of dimensions.

First, let a finite set of non-overlapping cells $(a_r^{(1)}, a_r^{(2)}; b_r^{(1)}, b_r^{(2)})$ $(r = 1, 2, 3, \ldots n)$ be contained in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, and let us suppose that $b_r^{(1)} - a_r^{(1)} \geqq \epsilon^{(1)}$, $b_r^{(2)} - a_r^{(2)} \geqq \epsilon^{(2)}$, for $r = 1, 2, 3, \ldots n$; where $\epsilon^{(1)}$, $\epsilon^{(2)}$ denote fixed positive numbers. The number n of the cells cannot exceed $(b^{(1)} - a^{(1)})(b^{(2)} - a^{(2)})/\epsilon^{(1)}\epsilon^{(2)}$.

For, let the fundamental cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ be divided into parts by constructing straight lines parallel to the $x^{(2)}$ axis through the points on the $x^{(1)}$ axis which are represented by the $2n$ numbers $a_1^{(1)}, b_1^{(1)}, a_2^{(1)}, b_2^{(1)}, \ldots a_n^{(1)}, b_n^{(1)}$, and by constructing straight lines parallel to the $x^{(1)}$ axis through those points of the $x^{(2)}$ axis that are represented by the $2n$ numbers $a_1^{(2)}, b_1^{(2)}, a_2^{(2)}, b_2^{(2)}, \ldots a_n^{(2)}, b_n^{(2)}$.

The sum of the products of the sides of these parts into which the fundamental rectangle is divided is clearly equal to $(b^{(1)} - a^{(1)})(b^{(2)} - a^{(2)})$. Every one of the rectangles of the given non-overlapping set is either identical with one of these parts, or is the sum of two or more of the parts. It then easily follows that

$$\sum_{r=1}^{r=n} (b_r^{(1)} - a_r^{(1)})(b_r^{(2)} - a_r^{(2)}) \leqq (b^{(1)} - a^{(1)})(b^{(2)} - a^{(2)});$$

and from this it follows that the number of the rectangles of the given set cannot exceed $(b^{(1)} - a^{(1)})(b^{(2)} - a^{(2)})/\epsilon^{(1)}\epsilon^{(2)}$.

Now let us consider a given non-finite set of cells, no two of which overlap, and all in the rectangle $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. Give to $\epsilon^{(1)}$ the values in a sequence of diminishing numbers $\epsilon_1^{(1)}, \epsilon_2^{(1)}, \ldots \epsilon_n^{(1)}, \ldots$ which converges to zero; $\epsilon^{(2)}$ having the values in another such sequence $\epsilon_1^{(2)}, \epsilon_2^{(2)}, \ldots$.

The number of cells of the given set which are such that one side is $\geqq \epsilon_n^{(1)}$, and the other $\geqq \epsilon_n^{(2)}$, is finite; for it is not greater than

$$(b^{(1)} - a^{(1)})(b^{(2)} - a^{(2)})/\epsilon_n^{(1)}\epsilon_n^{(2)}.$$

---

* Cantor, *Math. Annalen*, vol. xx (1882), p. 118 *et seq.*

We can now arrange the cells of the given set in order, taking first those for which the sides are $\geq \epsilon_1^{(1)}$ and $\geq \epsilon_1^{(2)}$; then those for which the sides are $< \epsilon_1^{(1)}, \epsilon_1^{(2)}$ and $\geq \epsilon_2^{(1)}, \epsilon_2^{(2)}$; and so on; there being only a finite number in each set. In each of such finite sets the cells are arranged in descending order of magnitude of the products of the side, or, in case of equality, in order of rank. Thus, since the whole set can be arranged as a simply infinite aggregate, the set is enumerable.

The case in which the cells of the set are in unbounded plane space can be considered as before by employing the correspondence

$$\xi^{(1)} = x^{(1)}/\{(x^{(1)})^2 + 1\}^{\frac{1}{2}}, \quad \xi^{(2)} = x^{(2)}/\{(x^{(2)})^2 + 1\}^{\frac{1}{2}}.$$

We have now the general result:

Any set of non-overlapping cells in p-dimensional space forms an enumerable aggregate.

It will be observed that, although geometrical language is employed above, the proof of the theorem is essentially arithmetical.

70. *Every isolated set is enumerable.*

This theorem may be proved by applying the theorem in § 69.

Let $P(x)$ be a point of an isolated linear set. The positive numbers h can be divided into two classes, the first consisting of the numbers h such that the closed interval $(x, x + h)$ contains no points of the given set, except the point x, and the second class consisting of those numbers h for which this is not the case. The two classes will be separated by a number h_1, where $x + h_1$ is either a point of the given set, or else is a limiting point of the given set but does not belong to it. A similar interval $(x - h_2, x)$ may be defined on the left of x; the interval $(x - h_2, x + h_1)$ contains no point of the given set in its interior, except x. Let ρ be the smaller of the two numbers $\frac{1}{2}h_2$, $\frac{1}{2}h_1$, and let the interval $(x - \rho, x + \rho)$ correspond to the point x. Taking the set of all intervals $(x - \rho, x + \rho)$ that correspond to the points of the given set, we see that this set of intervals is non-overlapping, and is consequently enumerable. Since each interval corresponds to the point x at its centre, it follows that the given isolated set of points is enumerable.

In the case of an isolated set in space of two, three, or more, dimensions, it can be shewn that, with each point x as centre, a circle, sphere, or hyper-sphere of radius 2ρ can be so determined that no point other than x is interior to it, and such that this is not the case for any circle, or sphere, of greater radius. We then take the circle, or sphere, of radius ρ, corresponding to each point x of the set. The set of all such circles, or spheres, will be non-overlapping. In every sphere a cell with equal edges can be inscribed; and the set of all such cells will be non-overlapping; to each cell there corresponds the point of the isolated set at its centre. It now follows, as before, that the given isolated set is enumerable.

Any set of points G is the sum of an isolated set and of one which is a component of G'. It follows that, if the derivative G' is enumerable, so also is G, but the converse does not hold.

Every set of points which is of the first species is enumerable.

For, if s be the order of G, the set $G^{(s)}$ contains only a finite set of points; and therefore $G^{(s-1)}$ is enumerable. Consequently $G^{(s-2)}, G^{(s-3)}, \ldots G$ are all enumerable sets.

A set of points of the second species is enumerable if one of its transfinite derivatives is enumerable. An unenumerable set G cannot have any enumerable derivative.

71. Let us consider a given set of overlapping intervals, not necessarily enumerable, contained in a finite segment (a, b) Let S be the set of those points of (a, b) each of which is an interior point of one interval at least of the given set.

If P be a point of S, it is interior to an interval α of the given set, and therefore a neighbourhood of P exists, such that every point of it is an interior point of α; and therefore all the points of that neighbourhood of P belong to S. Hence P is an interior point of S; and therefore S is an open set. Now it may be shewn that:

Every open linear set of points consists of the interior points of a finite, or enumerable, non-overlapping set of intervals.

For, if x be a point of S, it can be shewn, in exactly the same manner as in § 70, that an interval $(x - h_2, x + h_1)$ can be defined which has the property that every interior point of it belongs to S, but that $x - h_2, x + h_1$ do not belong to S. In this manner, to each point x, of S, there is correlated a definite interval $\delta(x)$; to all the points interior to $\delta(x)$ the same interval $\delta(x)$ corresponds. Now let us consider the set of all such intervals δ; these intervals are non-overlapping, and thus form a finite, or enumerable, set of intervals. Their interior points are identical with the set S.

The following theorem has been established:

Every set of intervals contained in a finite segment can be replaced by a set of non-overlapping intervals of which the interior points are the same as those of the given set.

If we consider the given set as a set of open intervals, the theorem is equivalent to the statement that the set of points, each of which belongs to one or more open intervals of the given set, is itself an open set. This is a generalization, for the case of an enumerable set of open sets, of a theorem given in § 56.

The new set may be spoken of as the set of non-overlapping open intervals equivalent to the given set of open intervals.

In the case of a p-dimensional set, it is shewn in exactly the same manner as in the case $p = 1$, that, having given a set of overlapping cells, the set S, of points, each of which is interior to one at least of the cells, is an open set. The nature of an open set in space of more than one dimension will be considered later.

72. The properties of any set of open intervals in a finite segment have been shewn to depend upon those of a non overlapping set of such intervals, and we therefore proceed to the consideration of the latter.

Every point of (a, b) which is not interior to an interval of the non-overlapping set is either

(1), a common end-point of two intervals of the given set, or

(2), a point interior to, or at an end of, an interval not belonging to the given set, this interval containing no point which is interior to any interval of the set; or

(3), a limiting point, on both sides, of end-points of intervals of the set; or

(4), an end-point of an interval of the given set, and also a limiting point, on one side, of end-points of intervals of the given set

If either a or b is an end-point of an interval, we reckon that point as belonging to the points (1).

The points described in (2) and (3) may be described as *external points* of the given set; and if a or b is a limiting point of end-points, it will be reckoned as an external point.

The points described in (4) may be spoken of as *semi-external** points.

Those points of the segment (a, b) *which are not points of a given set of non-overlapping open intervals form a closed set.*

This theorem is that particular case of the theorem proved in § 56, that the complementary set of an open set (relatively to a closed interval in which it is contained) is a closed set, which arises when the open set is the given set of non-overlapping open intervals.

It will now be shewn that†, *unless a given set of non-overlapping intervals is a finite set, there must be at least one external or semi-external point;* in other words, the whole interval (a, b) cannot be filled up by an indefinitely great number of non-overlapping intervals, each one of which abuts on the next, without leaving at least one point over, which is neither interior to an interval nor is an end-point of two intervals; the points a, b being regarded as end-points of two intervals if they are end-points of one interval of the given set.

---

* This term is due to W. H. Young, *Proc. Lond. Math. Soc.* (1), vol. xxxv (1902), p. 250.

† This theorem was given by W. H. Young, *ibid.* p. 251.

If there be any complementary intervals, *i.e.* intervals in which no point belongs to the given set of non-overlapping intervals, then the points of these intervals are all external points, and we therefore need only consider the case in which no such complementary intervals exist. We observe that, when the number of intervals is not finite, their end-points must have at least one limiting point P. Now this point P cannot be interior to one of the given intervals; for, if it were so, it would have a neighbourhood, viz. the interval to which it is interior, within which are no end-points. Neither can P be a common end-point of two intervals; for it would then have a neighbourhood on the right, and also one on the left, within which there is no end-point except P itself. The point P must consequently either be an external point, *i.e.* one which is not an end-point but is a limiting point, on both sides, of end-points; or else it must be an end-point of one interval, and a limiting point, on one side, of end-points. If a, or b, is not an end-point, it is regarded as an external point. It will subsequently be shewn that the external and semi-external points form a set which may be either finite, or of cardinal number a, or of cardinal number c.

Having given an overlapping system of intervals, those points that are end-points of intervals of the system and are not interior points of any interval of the system form a non-dense enumerable set.

If P be a left-hand end-point of an interval Pp of the given set, and if P is not interior to any interval of the set, Pp cannot have in its interior any other left-hand end-point Q of an interval Qq of the given system, such that Q is not interior to any interval of the set. Hence all such intervals Pp are non-overlapping, and therefore form an enumerable set.

Their left-hand end-points form an enumerable set, which is clearly non-dense. Similar reasoning applies to those points which are right-hand end-points of intervals of the given set, and are not interior to any interval.

Thus the theorem has been proved.

EXAMPLES

1*. In the interval $(0, 1)$ take the intervals $(0, \frac{1}{4})$, $(\frac{1}{4}, \frac{3}{8})$. ... $\left(\frac{2^{n-1}-1}{2^n}, \frac{2^n-1}{2^{n+1}}\right)$ ..., and also the intervals obtained by reflecting these intervals in the point $\frac{1}{2}$. The point $\frac{1}{2}$ is external to all the intervals, and yet the limiting sum of the intervals is equal to 1, the length of the whole interval $(0, 1)$ in which the enumerable set of intervals is contained.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (1), vol. xxxv (1902), pp. 249–251.

If instead of reflecting the intervals in the point $\frac{1}{2}$, we take the interval $(\frac{1}{2}, 1)$, the point $\frac{1}{2}$ is now a semi-external point, and the limiting sum of the intervals is the same as before.

2*. Take the set $(\frac{1}{2}, 1)$, $(0, \frac{1}{4})$, ... $\left(\frac{2^{n-1}-1}{2^n}, \frac{2^n-1}{2^{n+1}}\right)$ ... of intervals, and divide each interval into a set of sub-intervals similar to the whole. We now have a new enumerable set of intervals which has no external points, but of which the semi-external points form an enumerable set $\frac{3}{4}, \frac{1}{2}, \frac{1}{8}, \frac{5}{16}, \dots$.

THE HEINE-BOREL THEOREM

73. *If a set Δ of intervals, all in the fundamental interval (a, b), be such that every point of the closed interval (a, b) is an interior point of at least one interval of the set Δ (the end-points a, b being regarded as interior to an interval when either of them is an end-point of such interval), then a finite set of intervals, all belonging to Δ, exists, which has the same property as the set Δ, viz. that every point of the closed interval (a, b) is interior to at least one interval of the finite set.*

This theorem†, which is known as the Heine-Borel theorem, is of great importance in the Theory of Functions, and may be proved as follows:

Apply a system of nets with closed meshes to the interval (a, b). There must exist some value of n such that every mesh of the net D_n is interior to one interval at least of the set Δ. For, let us suppose, if possible, that this is not the case. Let α_n denote those meshes of D_n each of which is not interior to any interval of the set Δ. By hypothesis, α_n exists for every value of n; and it is clear that the closed sets $\alpha_1, \alpha_2, \dots \alpha_n, \dots$ are such that each set contains the next. In accordance with the theorem established in § 67, there is at least one point x that is contained in all the closed sets $\{\alpha_n\}$. The set of all such points x is closed; let $\bar{x}$ be the lower extreme of the set. The point $\bar{x}$ is interior to some interval $(\bar{x} - \epsilon, \bar{x} + \epsilon')$ of Δ; or to an interval $(a, a + \epsilon)$ in case $\bar{x} = a$; or to an interval $(b - \epsilon', b)$ in case $\bar{x} = b$.

When n is sufficiently large, d_n the maximum breadth of the meshes of D_n is less than both ϵ and ϵ'; or, when $\bar{x} = a$, is less than ϵ'; or when $\bar{x} = b$, is less than ϵ. For such a value of n, $\bar{x}$ is either interior to a mesh of D_n, which mesh is interior to the interval $(\bar{x} - \epsilon, \bar{x} + \epsilon')$, or else $\bar{x}$ is a common end-point of two meshes of D_n both of which are interior to that interval; or $\bar{x}$ may be at a or at b. In any case a mesh of α_n is interior to an interval of Δ; contrary to the supposition made above.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (1), vol. XXXV (1902), pp. 249–251.

† The first statement of the theorem was given by Borel, *Annales sc. éc. normale* (3), vol. XII (1895), p. 51. See also W. H. Young, *Proc. Lond. Math. Soc.* (1), vol. XXXV (1903), p. 301 and *Rendiconti di Palermo*, vol. XXI (1906), p. 125. The most general statement is that given by Gross, *Wiener Sitzungsber.* vol. CXXIII II A (1914), p. 809.

It has thus been shewn that, for a sufficiently large value of n, each of the meshes of D_n is interior to one at least of the intervals Δ; if, with each mesh of D_n there be associated an interval of Δ to which it is interior, the finite set of such intervals of Δ is a set such as is required.

In the above theorem it is not assumed that the given set of intervals Δ is enumerable. It has been proved that the interval (a, b) can be divided into a finite number of parts, each of which is interior to a set of intervals belonging to Δ. Each such set of intervals may be finite, or infinite (enumerable or not). The effective selection of a finite set of intervals of Δ, which satisfies the condition of the theorem, presupposes a knowledge of the mode in which the set of intervals Δ is defined. The theorem does not by itself provide any general method of making such effective selection. In the most general case, the interval (a, b) is divided into a finite number of parts, each of which has associated with it an (in general) infinite set of intervals of Δ, to all of which it is interior; the effective selection of the required finite set of intervals involves the choice of one of the possibly infinite set of intervals all of which contain one of these finite parts of (a, b). The number of such choices to be made, in each particular case, being finite, the principle known as the "multiplicative axiom," which will be discussed in Chapter IV, is not required in its most general form.

It is clear that the above proof is applicable to the case in which cells in any number of dimensions take the place of linear intervals The method of employment of systems of nets is precisely the same as in the case of the linear system. Accordingly we have the Heine-Borel theorem for the case of p-dimensional cells.

If a set of p-dimensional cells $\Delta^{(p)}$, *all in the fundamental cell* (a, b), *be such that every point of the closed cell* (a, b) *is an interior point of at least one cell of* $\Delta^{(p)}$ *(a point on the boundary of* (a, b) *being regarded as interior to a cell to whose boundary it belongs), then a finite set of cells, all belonging to* $\Delta^{(p)}$, *exists, which has the same property as the set* $\Delta^{(p)}$.

74. The Heine-Borel theorem can be generalized so as to apply to the case in which the points that are to be interior to a finite number of intervals, or cells, form any given closed set G, instead of consisting of all the points of the closed cell (or interval) in which the given set $\Delta^{(p)}$ is contained. The generalized theorem may be stated as follows:

Having given a (bounded) closed set G, *and a set of cells* $\Delta^{(p)}$, *such that each point of* G *is interior to one cell at least of the set* $\Delta^{(p)}$, *there exists a finite set of cells, all belonging to* $\Delta^{(p)}$, *such that each point of* G *is interior to one at least of the cells of the finite set.*

To establish this theorem a slight modification only of the proof of the theorem when G consists of all the points of (a, b) is required. We may choose the cell (a, b) so that it contains G, and we can neglect any part of the set $\Delta^{(p)}$ that is not in (a, b). Those meshes of the nets $D_1, D_2, \ldots$ which contain no points of G are throughout disregarded. The finite set α_n consists of those meshes of D_n that contain points of G and are not interior to any cell of $\Delta^{(p)}$. It is then clear that the point x is necessarily a point of the closed set G.

If the given set G were not closed, the point x might be a limiting point of G which did not belong to G, and thus the argument would break down.

The following theorem may be established by employing the generalized Heine-Borel theorem·

Having given a closed set, an enumerable set can be determined which has for derivative the given set.

With each point of the closed set G as centre let there be taken a cell of span d. A finite number of these cells are such that every point of G is interior to one of them; let this set be C_d, and let $P_1^{(d)}, P_2^{(d)}, \ldots P_{n_d}^{(d)}$ be their centres, all points of G. Let d have a sequence of decreasing values $d_1, d_2, \ldots$ converging to zero.

The totality of the points $P_1^{(d)}, \ldots P_{n_d}^{(d)}$, taken for every value of d in the sequence, is an enumerable set H of points of G. Since H consists of points of the closed set G, every limiting point of H belongs to G Every point of G which does not belong to H is interior to all the cells of a sequence of cells of spans $d_1, d_2, \ldots$ and is therefore a limiting point of the centres of the cells of this sequence; therefore a point of G which does not belong to H belongs to H'. Considering those points of G which belong to H, but not to H', we see that these are all isolated points of G, and each one of them is the centre of some cell which contains no point of G other than its centre. In each such cell place an enumerable set of points having the centre of the cell for its sole limiting point. Let the totality of all these enumerable sets be the set K, which is enumerable, since the isolated points of G form an enumerable set. The set K' consists of the points of G which belong to H but not to H', together with the limiting points of the centres of the cells employed, and these also belong to G. The enumerable set $H + K$ has G for its derivative.

75. It has been pointed out* by de la Vallée Poussin that the generalized form of the Heine-Borel theorem still holds good in case $\Delta^{(p)}$ is replaced by an aggregate S of open sets. More generally, instead of $\Delta^{(p)}$, any aggregate of sets may be taken, such that each point of G is an interior point of one or more of the sets of the aggregate. This may be stated as follows:

---

* See *Intégrales de Lebesgue*, pp. 15, 112.

If a closed (bounded) set G and an aggregate S of sets be such that every point of G is an interior point of one at least of the sets of S, a finite number of the sets of S exists such that each point of G is an interior point of at least one of them.

To prove this theorem, it may be observed that, with each point P of G a cell may be associated, with P in its interior, and such that the cell is interior to one or more of the sets of S. The aggregate of all such cells may be taken to be the set $\Delta^{(p)}$ in the generalized form of the Heine-Borel theorem. A finite part of the set $\Delta^{(p)}$ exists such that each point of G is interior to one or more of the cells of this finite set. Each of these cells is contained in a set belonging to S. Hence there exists a finite aggregate of the sets of S which has the required property.

In case the sets S form an aggregate $H_1, H_2, \ldots H_\omega, \ldots H_\beta, \ldots$ such that each one is contained in the next, the closed set G, each point of which is in one or more of the open sets, must be contained in H_β, where β is some fixed number of the first or of the second class.

For a finite number of the sets H exists such that every point of G is contained in one or more of them. If β be the index of that one of this finite set which has the highest index, G must be contained in H_β.

From this theorem, the following may be deduced.

If $G_1, G_2, \ldots G_\omega, \ldots G_\beta, \ldots$ are closed sets such that each contains the next, and F is a closed set of points such that no point of F is contained in all the sets G, there exists a definite number β, of the first, or of the second, class, such that no point of F belongs to G_β.

To prove this, we consider, with respect to some cell containing G_1, open sets $C(G_1)$, $C(G_2)$, complementary to $G_1, G_2, \ldots$. Each of these open sets is contained in the next. Any point of F is contained in some of the sets $C(G_1)$, $C(G_2)$, ...; and the set F is therefore contained in $C(G_\beta)$ for some value of β, and is therefore not contained in G_β.

It has already been pointed out that, when G is a set that is not closed, and each point of G is interior to at least one of the cells of a given set, there does not necessarily exist any finite part of the given set of cells such that every point of G is interior to one of them. However, in any case, an enumerable part of the given set of cells exists which has this property. Thus:

If G be any set of points, not necessarily closed, and each point of G be interior to one at least of the cells of a given unenumerable set $\Delta^{(p)}$, there exists an enumerable set of cells all belonging to $\Delta^{(p)}$, such that every point of G is interior to one at least of the cells of this enumerable set.

This theorem can, as in the case of the Heine-Borel theorem, be proved by means of a system of nets applied to the finite, or infinite, interval, or cell, in which G is contained. Moreover*, instead of $\Delta^{(\nu)}$, an aggregate of sets can be substituted, such that every point of G is an interior point of at least one of the sets of the aggregate.

Of a net D_n, for any large enough value of n, say n_1, some of the meshes containing points of G will be interior to one or more of the cells (or intervals) of the given unenumerable set Let these meshes be denoted by β_1. For values of n greater than n_1, we consider only those meshes of D_n that are not contained in β_1. For some such value of n, say n_2, there will be meshes β_2, containing points of G, not contained in β_1, each of which is contained in one or more of the intervals, or cells. of the given set. We obtain in this way a sequence $\beta_1, \beta_2, \beta_3, \ldots$ of finite (or enumerable, in case G is unbounded) sets of meshes, each of which contains points of G, and is contained in one or more of the intervals, or cells, of the given set. Each point of G must be contained in one of the meshes belonging to this set $\{\beta\}$. The meshes of $\beta_1, \beta_2, \ldots$ form an enumerable set, and corresponding to each of such meshes, there exists a cell (or interval) of the given set. Hence there exists an enumerable part of the given set such that every point of G is interior to one or more of the intervals, or cells, of this enumerable part.

In the choice of one out of a possibly infinite number of cells which have a given mesh in their interiors, the number of choices being in this case indefinitely great, the "multiplicative axiom" is in general involved

However, the following theorem may be established†, which is sufficient for applications, and in the proof of which no use is made of the multiplicative axiom:

If each point of a given set is interior to a definite cell corresponding to that point, an enumerable set of these cells can be so determined that all the points of the set are interior to one or more of the cells of the enumerable set.

If the set G be not bounded, we may place it in correspondence with a bounded set, so that, to the given cells, there corresponds a set of cells having the corresponding property with reference to the bounded set of points. There is accordingly no loss of generality in considering a bounded set only, which we can suppose to be contained in a fixed cell with equal sides. Instead of the cell of the given set $\Delta(P)$ corresponding to a point P of the given set G, we may substitute a cell $\Delta'(P)$ with the point P as centre and with its sides all equal, and take $\Delta'(P)$ to be the greatest such cell which has no point exterior to $\Delta(P)$. We have now, corresponding to each point P of the given set, a cell $\Delta'(P)$ with P as centre. Take a set of positive diminishing numbers $\{\epsilon_n\}$ which converges to zero, and let G_n be that part of G which consists of points for which the spans of the corresponding cells Δ' are $> \epsilon_n$. Since every point of G_n is in a cell (with equal sides) of spans all $> \epsilon_n$, every point of G_n' is in a cell of this system. Hence a finite set of these cells can be determined which contain in their interiors all the points of G_n', and therefore all the points of G_n with the possible exception of a finite number. By adding the cells corresponding to this finite number of points of G_n, we obtain a finite set of the cells Δ' which contain all the points of G_n. We have now a finite set of cells which contain all the points of G_1, and a finite set which contain all the points of G_2 that do not belong to G_1; and generally, a finite set of cells which contain all the points of G_n that do not belong to $G_1, G_2, \ldots$ or G_{n-1}. Taking every value of n, we have now an enumerable set of the cells Δ' that contain all the points of G, and we take the set of cells Δ which correspond to the enumerable set of cells Δ' as the required set.

---

* See Lindelöf, *Comptes Rendus*, vol. CXXXVII (1903), p. 697.

† See Borel, *Leçons sur les fonctions monogènes*, p. 12.

76. In a proof given by Heine* that every continuous function is uniformly continuous there is contained the germ of the theorem now called the Heine-Borel theorem, as it was first explicitly stated and proved by Borel†, for the case of a linear interval (a, b). Various proofs of the theorem have since been published‡. It was pointed out by H F. Baker§ that the proof by Goursat‖ of Cauchy's fundamental theorem in the theory of functions of a complex variable contains a general method of procedure, equivalent to the use of nets made in § 73. The extension to the case of any closed set G was given¶ by W. H. Young, and by Borel**.

By means of an analysis of the reasoning of Heine, W. H. Young was led†† to the discovery of the following more general theorem*, called the Heine-Young theorem:

If, with each point x of a closed interval (a, b), we have associated a pair of intervals r_x and l_x, such that, (1), x is the left-hand end-point of r_x and the right-hand end-point of l_x and, (2), if x' is an internal point of l_x, then x is an internal or end-point of $r_{x'}$; there then exists a finite number of the intervals r_x abutting end to end and covering the whole segment (a, b).

It is unnecessary to suppose r_b and l_a to be defined.

Starting with a as end-point, we take the interval r_a, or (a, x_1); we then take the interval r_{x_1} or (x_1, x_2); and so on. We thus obtain a series of intervals of the given set, $(a, x_1)\ (x_1, x_2) \ldots (x_{n-1}, x_n) \ldots$.

---

* *Journal f. reine und angewandte Mat.* vol. LXXIV (1871), p. 188.

† *Annales sc. éc. normale* (3), vol. XII (1895), p. 50.

‡ See for example Borel's *Leçons sur l'intégration*, p. 105.

§ *Proc. Lond. Math. Soc.* (1), vol. XXXV (1902), p. 459, and (2), vol. I (1903), p. 24.

‖ *Trans. Amer. Math Soc.* vol. I (1900), p 15.

¶ *Proc. Lond. Math. Soc.* (1), vol. XXXV (1902), p. 387. For a further extension of the theorem, see W. H. Young, *Mess. of Math.* vol. XXXIII (1904), p. 129, and also *Proc. Lond. Math. Soc.* (2), vol. II (1905), p. 67.

** *Comptes Rendus*, vol. CXL (1905), p. 298.

†† *Proc. Lond. Math. Soc.* (2), vol. XIV (1914), p. 114.

If, for some value of n, $x_n = b$, the condition of the theorem is satisfied by the n intervals thus obtained. Otherwise, the sequence $x_1, x_2, \ldots$ has a limiting point X on the right of all the points $\{x_n\}$. There is then an interval l_X with X as right-hand end-point.

All the points $\{x_n\}$, except a finite number of them, are interior to the interval l_X; let these interior points be $x_n, x_{n+1}, \ldots$. By the hypothesis (2) in the statement of the theorem, r_{x_n} must reach at least as far as X, and this is not the case, as it reaches only to x_{n+1}. Hence the sequence $\{x_n\}$ cannot be infinite. Thus the theorem is established.

In case there exists no interval l_b, so that the conditions of the theorem hold for the half-open interval (a, b), there exists an enumerable set of the intervals r_x abutting on one another and covering the half-open interval (a, b). For it is clear that in this case the point X must coincide with b.

The following theorem, originally due to Lusin*, and employed by him in the theory of derivatives of a function, follows from the Heine-Young theorem:

If, associated with every point x of a closed interval (a, b), we have all the intervals with x as end-point that lie in a certain neighbourhood of the point x on both sides (except for the points a, b where the neighbourhood is on one side), then a finite number of these intervals exists, abutting end to end, and covering the interval (a, b).

At each point x we take for r_x the smallest interval with x as left-hand end-point containing all those of the given intervals which have x for end-point, whether originally associated with x or not; and for l_x the part of the given neighbourhood on the left of x.

There is then a finite set of these intervals r_x covering (a, b), and abutting on one another. Let these be $(a, x_1)\ (x_1, x_2) \ldots (x_{n-1}, b)$; and consider (x_{r-1}, x_r). We may choose an interval of the given set with x_{r-1} as left-hand end-point; if this interval does not reach to x_r we add one of the given intervals associated with the point x_r, and thus two of the given intervals make up (x_{r-1}, x_r). Doing this for each interval (x_{r-1}, x_r), by a finite number of choices, we obtain the required finite set which covers (a, b).

By means of Lusin's theorem the Heine-Borel theorem can now be deduced. If we replace those intervals of Δ which contain x as interior point by all intervals with x as end-point contained in each one of the intervals of Δ, we have the condition in Lusin's theorem satisfied. The finite set of abutting intervals which covers (a, b) having been determined, we replace each one of the intervals of that finite set by an interval of Δ that contains it; this requires only a finite number of choices. Thus the Heine-Borel theorem is deduced.

---

* *Recueil de la soc. mat. de Moscou*, vol. XXVIII, 2 (1911).

77. The following theorem will now be proved:

If any unenumerable set of overlapping intervals in (a, b) *be given, then an enumerable set of intervals all belonging to the given set exists, of which the interior points are the same as those of the given set.*

It has been shewn in § 71 that the given set can be replaced by a non-overlapping set of intervals with the same interior points. An interval of this second set is however not in general an interval of the given set.

Let PQ be an interval of the equivalent non-overlapping set; then every internal point of PQ is an internal point of one interval at least of the given set. The point P is either an end-point of some interval Pp of the given set, or else it is a limiting point of end-points of an infinite number of intervals of the given set. In the latter case there exists an enumerable sequence $P_1p_1, P_2p_2, P_3p_3, \ldots$ of intervals of the given set such that P is the limiting point of the sequence of points $P_1, P_2, \ldots P_n, \ldots$ Similarly, unless Q is an end-point of an interval qQ of the given set, it is the limiting point of a sequence $Q_1, Q_2, \ldots Q_n, \ldots$ of end-points of intervals q_1Q_1, $q_2Q_2, \ldots q_nQ_n, \ldots$ of the given set. Consider the intervals P_1Q_1, $P_2Q_2, \ldots$ $P_nQ_n, \ldots$, where $P_1, P_2, \ldots$ may be taken all to coincide with P in case the interval Pp belongs to the given set, a similar convention being made as regards Q. Since every point of P_1Q_1 is interior to some interval of the given set, therefore, in accordance with the Heine-Borel theorem, a finite number of intervals of the given set exists, such that every point of P_1Q_1 is interior to one at least of them. Let a similar determination of a finite set of intervals be made for each of the intervals $P_2Q_2, P_3Q_3, \ldots P_nQ_n, \ldots$: we have then altogether an enumerable set of finite sets of intervals. The totality of these intervals forms a finite, or an enumerable, set of intervals belonging to the given set, which contains every point in the interior of PQ as an interior point. Applying the same process to each interval PQ of the equivalent non-overlapping set, and remembering both that the intervals PQ form a finite, or an enumerable set, and that an enumerable set of finite or enumerable sets is itself enumerable, we derive the conclusion that an enumerable set of intervals exists, all belonging to the given set, such that their internal points are identical with those of the given set.

The assumption, here made, that, if P be a limiting point of end-points of intervals of the given set, a sequence of such end-points exists which converges to P, will be discussed in Chap. IV.

THE LEBESGUE CHAIN OF INTERVALS

78. For the purposes of the Theory of Functions it is convenient to consider a set of intervals in a given linear segment, of the kind known as a Lebesgue chain.

If (a, b) *be a given linear interval, a set* C_a^b *of non-overlapping closed intervals such that every point of the closed interval* $(a \quad x \quad b)$ *is either a left-hand end-point of one of the intervals or an internal point of one of these intervals, is said to form a chain stretching from* a *to* b. *The chain is supposed to have a last interval which contains* b *as interior point or as left-hand end-point.*

To consider the nature of such a chain, let us suppose that we have defined an interval (a, x_1) with a as left-hand end-point, then a set of intervals (x_1, x_2), (x_2, x_3), ... from left to right. In case the right hand end-point of one of these intervals is b, we have, by adding one interval, a chain consisting of a finite number of closed intervals, reaching from a to b. But, if this is not the case, the points $x_1, x_2, \ldots x_n, \ldots$ will have a limiting point ζ_1 on their right, which may coincide with b, in which case we have, by adding one more interval, a chain reaching from a to b consisting of an enumerable set of intervals. But if $\zeta_1 < b$; starting from ζ_1, a set of intervals (ζ_1, ζ_2), (ζ_2, ζ_3), ... is supposed to be defined. If these reach from ζ_1 to b; either by taking a finite number, or by taking an enumerable set, together with one more interval, δ, we have again a chain from a to b, consisting of the intervals

$$(x_1, x_2), (x_2, x_3) \ldots (\zeta_1, \zeta_2), (\zeta_2, \zeta_3) \ldots, \delta.$$

But if $\zeta_1, \zeta_2, \ldots$ have a limiting point η_1 $(< b)$, we proceed as before to set up new intervals proceeding from η_1. This process may be carried on indefinitely, until a chain is defined reaching from a to b. That this will always happen when any set of rules is prescribed for the definition of the intervals is asserted in the following theorem*:

If, for each point of the closed interval (a, b), *there be assigned, by some prescribed set of rules, one single interval with the point as left-hand end-point, then, if* X *be any point in the closed interval* (a, b), *there is one and only one chain stretching from* a *to* X, *composed of intervals of the given set.*

It will be observed that the theorem asserts (1), the existence of a chain from a to X, and (2), its uniqueness. Also, if a chain stretching from a to X exists, and X' be any point on the left of X, a chain from a to X' exists which is a part, or the whole, of the chain from a to X.

---

* See Pal, *Rend. di Palermo*, vol. XXXIII (1912), p. 352; also G. C. Young, *Quart. Journ. of Math.* vol. XLVII (1916), p. 142, and W. H. & G. C. Young, *Proc. Lond. Math. Soc.* (2), vol. XIV (1915), p. 128

If (a, x_1) be the interval corresponding to a, the theorem is clearly true for all points x such that $a \leq x \leq x_1$.

Let us assume, if possible, that there exists in (a, b) a set of points G, for which the property in the theorem does not hold good. The set G must have a lower boundary point X $(> x_1)$, which may or may not belong to G. Let us first suppose that X does not belong to G; it is consequently a limiting point of G. There is, by hypothesis, a unique chain $C_a{}^X$ reaching from a to X. Since $C_a{}^X$ has a last interval $(X - \epsilon', X + \epsilon)$, or $(X, X + \epsilon)$, the chain $C_a{}^X$ reaches from a to ζ, where ζ is such that $0 < \zeta - X < \epsilon$. Now $C_a{}^X$ is the only chain that can reach from a to ζ; for any chain that reaches from a to ζ must contain some chain that reaches from a to X, and this can only be the chain $C_a{}^X$. It follows that, in a certain neighbourhood of X on its right, there are no points of G, which is contrary to the hypothesis that X is a limiting point of G.

Next, let us suppose that X belongs to G, whether it be a limiting point of G or not. Consider a sequence of points $x_1, x_2, \ldots x_n, \ldots$ of which X is the limiting point, and such that $x_1 < x_2 < x_3 < \ldots < x_n < \ldots$. A unique chain $C_a{}^{x_n}$ stretches from a to x_n; also if $n' > n$, $C_a{}^{x_n}$ is a part, or the whole, of $C_a{}^{x_{n'}}$. Consider the complete set C of intervals each of which belongs to $C_a{}^{x_n}$, from and after some value of n. Then C, together with the interval $(X, X + \epsilon)$, corresponding to the point X, constitutes a chain reaching from a to X. For each point on the left of X is reached by $C_a{}^{x_n}$ for some value of n, and is therefore an interior point, or a left-hand end-point of an interval of C. Any chain reaching from a to X contains as a part a chain from a to x_n, and this can only be $C_a{}^{x_n}$. As this holds for every value of n, any chain from a to X must contain $C_a{}^{x_n}$ for every value of n; and therefore it contains C. Also C must be independent of the particular sequence $\{x_n\}$; for any point x'_m of another sequence must lie in an interval (x_n, x_{n+1}), and the unique chain from a to x'_m must be a part, or the whole, of the chain from a to x_{n+1}. Any chain from a to X can thus only consist of C and the interval $(X, X + \epsilon)$; thus there is a unique chain reaching from a to X, contrary to the hypothesis that X belongs to G. It now follows that no such set as G can exist, and the theorem is therefore established.

79. The reasoning by which Lebesgue established the existence of a chain from a to b depends upon the employment of transfinite numbers of the second class, and upon the principle that every enumerable ordered aggregate must be exhausted before some particular number of the second class is reached. If a set of intervals

$$(a, x_1)\ (x_1, x_2) \ldots (x_{n-1}, x_n) \ldots (x_\omega, x_{\omega+1}) \ldots (x_\beta, x_{\beta+1}) \ldots$$

be numbered from left to right, the numbers employed will all be of the first, or of the second, class. If the process of construction does not cease before some particular number of the second class is used, the set of intervals would be unenumerable; and, as they are non-overlapping, this is impossible (§ 69). Thus, by a set of such intervals, starting with a, the point b must be reached, and the existence of a chain reaching from a to b is established.

It may happen however that, corresponding to each point x of (a, b), not a unique interval, but a finite or infinite set of intervals is defined, with x as left-hand end-point. In that case, the theorem as to the existence of a Lebesgue chain takes a more general form than that of Pal, given above, and the proof given in § 78 fails*. For chains from a to x_1 and from a to x_2, where $x_2 > x_1$, may both exist, but it does not follow that the former is part of the latter. The general existence theorem may be stated as follows:

If, to each point x of the interval (a, b) (except b) there be assigned a set of intervals (finite or infinite) with x as left-hand end-point, an enumerable non-overlapping set of the intervals can be so determined as to contain every point of the semi-open interval (a, b) as interior or left-hand end-point.

A proof of this theorem not involving an infinite number of acts of choice is a desideratum.

For most of the purposes of the theory of functions the more restricted form of the theorem suffices, in which a single interval is defined, with each point of (a, b) as left-hand point.

CLOSED AND PERFECT LINEAR SETS

80. The following theorem may be stated:

Those points of the linear segment (a, b) which are not points of a given set of non-overlapping open intervals form a closed set of points.

This theorem has been proved in § 56, where it was shewn that the complementary set of an open set, relatively to a closed interval in which it is contained, is a closed set. In case a, or b, is a point of the set of open intervals, it is regarded as an interior point.

Conversely, it has been shewn in § 56, that:

Every closed set of points in the linear segment (a, b) is the complement of a non-overlapping set of open intervals.

In accordance with the classification given in § 72 of the points which do not belong to a set of open intervals, it appears that:

The most general linear closed set of points in an interval (a, b) consists of (1), the end-points of a set of non-overlapping intervals, (2), limiting points of such end-points, and (3), the points interior to intervals every point of which belongs to the closed set.

---

* W. H. & G. C. Young, *Proc. Lond. Math. Soc.* (2), vol. XIV (1915), pp. 128–130.

The open intervals belonging to the set $C(G)$, complementary, relatively to (a, b), to a given closed set G in (a, b) are said to be the intervals *contiguous* to, or *complementary* to, the said G. The set of all such intervals may be spoken of as the set of *contiguous intervals*, or *complementary intervals*, for the closed set G.

If the set G is non-dense, no interval exists in (a, b) which consists entirely of points of G, and the set of contiguous intervals is then everywhere dense in (a, b), since no interval can be determined in (a, b) so as to contain no points of the set of contiguous intervals. Thus:

Every linear non-dense closed set in an interval* (a, b) *consists of the end-points of the intervals of an everywhere dense set of open intervals and of the limiting points of such end-points.*

The most general type of closed set is obtained by adding to a non-dense closed set all the points of some of the contiguous intervals.

The points of a non-dense closed set G consist in general of three classes:

(1) those which are common end-points of two contiguous intervals abutting on one another;

(2) semi-external points of the set of contiguous intervals, *i.e.* points which are end-points of one interval, and also limiting points, on one side, of end-points; and

(3) external points, *i.e.* such as are not end-points of any contiguous interval, but are limiting points, on both sides, of such end-points.

The point a, or the point b, if it belongs to G, may be regarded as belonging to (1) or (3), according as it is, or is not, an end-point of a contiguous interval.

Those points which belong to (1) are clearly isolated points of G. Hence, if no such points exist, every point of G is a limiting point: and therefore G is perfect. It follows that:

Every non-dense perfect linear set G consists of the end-points of an everywhere dense set of non-overlapping intervals (contiguous to G), no two of which abut on one another, together with the limiting points of these end-points.

If the closed set G is such that no semi-external points exist, then every contiguous interval abuts on another one at both its ends. In this case, all the points of G are either end-points of adjacent intervals, or limiting points, on both sides, of a sequence of such end-points; unless a or b be a limiting point, in which case it belongs to G. Such a set may be either enumerable, or it may contain a perfect set*, in which case it is unenumerable (see § 83). A closed set may be enumerable, and yet there may be contiguous intervals which do not abut on another contiguous interval. For example, let the contiguous intervals of a set G in the interval (0, 1) consist of the open intervals

---

* This relation between everywhere dense sets of intervals and closed sets was discovered by Du Bois-Reymond and by Harnack. See Du Bois-Reymond's *Allgemeine Functionentheorie* (1882), p. 188, also *Math. Annalen*, vol. XVI (1880), p. 128, where everywhere dense sets of intervals are introduced. See also Harnack, *Math. Annalen*, vol. XIX (1882), p. 239, and Bendixson, *Acta Math.* vol. II (1883), p. 416, and *Öfv. af Svensk. Vet. Förh.* vol. XXXIX (1883), part 2, p. 31. Proofs of the fundamental theorems based on the amalgamation of abutting intervals have been given by W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. I (1903), p. 240, and by Schoenflies, *Göttinger Nachrichten*, 1903.

$$\left(0, \frac{1}{2}\right) \text{ and } \left(\frac{1}{2} + \frac{1}{n+2}, \frac{1}{2} + \frac{1}{n+1}\right),$$

where $n = 1, 2, 3, \ldots$. The interval $\left(0, \frac{1}{2}\right)$ does not abut on another interval, so that $\frac{1}{2}$ is a semi-external point. This set G is enumerable.

To construct an unenumerable closed set which has no semi-external points, let H be a non-dense perfect set. In each interval (α, β) contiguous to H, place the enumerable set consisting of the points $\alpha + \frac{k}{n}$, $\beta - \frac{k}{n}$, where $n = 1, 2, 3, \ldots$, and k is a number less than $\frac{\beta - \alpha}{2}$. The totality of these sets for all the intervals (α, β) is an enumerable set K. The set $H + K$ is closed and unenumerable; each of its contiguous intervals abuts on another one at both its ends. A semi-external point α, of H, is an external point of the closed set $H + K$.

81. *Every non-dense linear closed set is, in general, made up of an enumerable set and of a perfect set.*

Let the intervals complementary to the set G be arranged in enumerable order, that of descending magnitude; we may denote them by $\delta_1, \delta_2, \ldots \delta_n, \ldots$. If G is not perfect, it contains isolated points, each of which is the common end-point of two adjacent intervals; let δ_{p_1} be the first of the intervals $\{\delta\}$ at an end of which there is such a point; let $\delta_{p'}$ be the interval which abuts on δ_{p_1} at that end. It may happen that the other end-point of $\delta_{p'}$ is also a common end-point of two intervals. If so, let $\delta_{p''}$ be the interval which abuts on $\delta_{p'}$, and so on: after a finite, or enumerable, set of such intervals $\delta_{p'}, \delta_{p''}, \delta_{p'''} \ldots$ we must arrive at an interval of which the end-point does not belong to G_ι, the set of isolated points of G, or else at an end-point of the domain of G. It may happen that δ_{p_1} at its other end abuts on another interval; in that case we proceed, in the same manner as before, to find the intervals $\delta_{q'}, \delta_{q''}, \ldots$ each of which abuts on another one. Now conceive all the intervals

---

* It was incorrectly stated by Schoenflies, in his *Bericht*, vol. I, p. 78, and also in the second edition of the present work (vol. I, p. 113) that a closed set is necessarily enumerable when each contiguous interval abuts on another one at both its ends.

$\delta_{p'}, \delta_{p''}, \ldots$, and, if they exist, $\delta_{q'}, \delta_{q''}, \ldots$ to be amalgamated with δ_{p_1} into one interval $\delta_{p_1}^{(1)}$, by removing all the common end-points. If any isolated points of G now remain, let δ_{p_2} be the first interval of $\{\delta\}$ after δ_{p_1}, of which an end-point is such a point; proceed as before, we then have an interval $\delta_{p_2}^{(1)}$ formed by amalgamating a finite or enumerable set of intervals. We proceed in this way, and thus form a set of intervals $\delta_{p_1}^{(1)}, \delta_{p_2}^{(1)}, \ldots$ no end-points of which are points of G_ι.

Since $G = G_\iota + G^{(1)}$, where $G^{(1)}$ is the derivative of G, the set of intervals $\{\delta^{(1)}\}$ complementary to $G^{(1)}$ consists of the intervals $\delta_{p_1}^{(1)}, \delta_{p_2}^{(1)}, \ldots$ and of any intervals $\{\delta\}$ which remain after such intervals as

$$\delta_{p'}, \delta_{p'}, \ldots \delta_{q'}, \delta_{q''}, \ldots$$

have been removed, and the $\delta_p^{(1)}$ substituted for the δ_p.

We proceed in a similar manner with $G^{(1)} \equiv G_\iota^{(1)} + G^{(2)}$, again removing a finite or enumerable number of the set $\{\delta^{(1)}\}$, and again with $G^{(2)}$, and so on. It may happen that the process comes to an end after a number n of such stages, either if $G_\iota^{(n)}$ does not exist, in which case $G^{(n)} = G^{(n+1)}$, and thus $G^{(n)}$ is perfect; or else, if $G^{(n)}$ does not exist, in which case G, being the sum of a finite number of enumerable sets $\Sigma G_\iota^{(r)}$, is itself enumerable. If the process does not come to an end for any finite value of n, we form the derivative $G^{(\omega)} = D\,(G^{(1)}, G^{(2)}, \ldots G^{(n)}, \ldots)$, which contains all the points common to all the derivatives of G of finite order. This set has been shewn, in § 68, to exist, and to be a closed set: $G^{(\omega)}$ is then resolved as before into $G_\iota^{(\omega)} + G^{(\omega+1)}$, and we proceed further as before.

We obtain, by proceeding in this manner,

$$G = G_\iota + G_\iota^{(1)} + \ldots + G_\iota^{(\omega)} + G_\iota^{(\omega+1)} + \ldots + G_\iota^{(\beta)} + G^{(\beta+1)},$$

where β is a number of the first or second class. It will now be shewn that there must be some definite number β of the first or second class, for which this process comes to an end, either by $G_\iota^{(\beta)}$ containing no points, in which case $G^{(\beta)} = G^{(\beta+1)}$, so that $G^{(\beta)}$ is perfect; or else by $G^{(\beta-1)}$ containing no points, in which case G, being the sum of an enumerable set of finite, or enumerable, sets, is itself enumerable. The $\{\delta\}$ contain all the indices 1, 2, 3, ... n, ...; from these indices we must remove a finite, or an enumerably infinite number, to obtain those indices which occur in the $\{\delta^{(1)}\}$; and again an enumerable set of indices must be removed from those which occur in the $\{\delta^{(1)}\}$, to obtain those which occur in the $\{\delta^{(2)}\}$. Now as the indices 1, 2, 3, ... n, ... are enumerable, the process of removing successively a finite, or enumerably infinite, set of them must cease for some order β of $\delta^{(\beta)}$, for otherwise a more than enumerable infinity of indices could be removed from the set 1, 2, 3, ... n, ..., which is impossible; hence, for some fixed number β of the second class, all the indices must have been removed.

It has thus been shewn that, unless the given set G is enumerable, for some number β of the first or second class, $G^{(\beta)}$ $G^{(\beta+1)}$; and therefore $G^{(\beta)}$ is perfect. Thus G has been resolved into an enumerable set and a perfect one, which may be called the *nucleus* of G.

If, for any value of β, $G^{(\beta)} \equiv 0$, the set G is enumerable.

82. The following theorem*, more general than that of § 81, includes the latter as a particular case. The proof here given may be taken as alternative to that of § 81.

If $P_1, P_2, \dots P_n, \dots P_\beta, \dots P_\alpha, \dots$ are all closed linear sets of points such that (1), *if $\alpha_1 < \alpha_2$, all the points of P_{α_2} belong to P_{α_1}, and* (2), *if in any interval, any set P_α contains only a finite number of points, the set $P_{\alpha+1}$ contains no points in that interval; then either P_β must vanish for some definite number β of the first or second class, or else there is a definite number β such that P_β is a perfect set.*

If, for some number β, the set P_β vanishes, then P_γ vanishes for all values of γ which are $> \beta$.

Let us now suppose that there exists no number β such that P_β vanishes. In this case there exists a set of points, which may be denoted by P_Ω, such that each point of the set belongs to P_β, whatever number β may be. The set P_Ω is closed, for if p be a limiting point of the set, in an arbitrarily small neighbourhood there are points of P_β, whatever number β may be; hence p belongs to P_β, whatever β may be, and thus p itself belongs to P_Ω.

It will now be shewn that P_Ω contains no isolated points, and is therefore dense in itself. If P_Ω contains an isolated point p, a neighbourhood of p can be found which contains no point of P_Ω except p; let Q be that part of P_1 which is contained in this neighbourhood. In the neighbourhood considered, let us suppose a sequence of intervals $\delta_1, \delta_2, \dots \delta_n, \dots$ constructed, each one containing the next one and the point p, and such that δ_n converges to zero as n is indefinitely increased. Let $Q^{(n)}$ denote that part of Q which lies in δ_n but not in δ_{n+1}, then

$$Q = Q^{(1)} + Q^{(2)} + \dots + Q^{(n)} + \dots + p.$$

There must exist a number β_1, of the first, or of the second, class, for which $Q^{(1)}$ contains no point of P_{β_1}; otherwise $Q^{(1)}$ would contain points which belong to P_Ω, and this is not the case. Similarly, there exist numbers $\beta_2, \beta_3, \dots \beta_n, \dots$ such that $Q^{(2)}$ contains no points of P_{β_2}, and $Q^{(3)}$ contains no points of P_{β_3}, and so on. Of the numbers $\beta_1, \beta_2, \dots \beta_n, \dots$, let γ_1 be the first which is $> \beta_1$, then let γ_2 be the first which is greater than γ_1, and so on; we have therefore a sequence $\gamma_1, \gamma_2, \dots \gamma_n, \dots$ of increasing numbers all of which belong to the set $\beta_1, \beta_2, \dots \beta_n, \dots$. This sequence $\gamma_1, \gamma_2, \dots \gamma_n, \dots$ is either finite, with say γ as the last, or else there is a limiting number γ of the second class which is greater than all of them, and therefore greater than all the numbers $\beta_1, \beta_2, \dots \beta_n, \dots$. The set Q can have no point, except p, which belongs to P_γ, hence, since P_γ contains only one point in a certain interval, $P_{\gamma+1}$ contains no point in that interval, and does not contain p; which is contrary to the hypothesis.

---

* See Baire, *Annali di Mat.* (3), vol. III (1899), p. 46 *et seq.*

It has now been shewn that P_Ω is closed and dense in itself; it is therefore perfect. Let us next consider the enumerable set of intervals which are complementary to P_Ω. For any one of these intervals there exists a number γ such that P_γ contains no point in the interior of the interval. As before, it is seen that there exists a number, of the first or the second class, which is greater than all these numbers γ; if this number be β, the set P_β contains no points which do not belong to P_Ω. It is thus seen that P_β is perfect, and $P_\beta \equiv P_{\beta+1} \equiv \dots \equiv P_\Omega$. The theorem has now been completely established.

If, in the above theorem, the condition (2) be omitted, the set P_Ω is closed but not necessarily perfect. The set P_β, as before, contains no points that do not belong to P_Ω; it then follows that

$$P_\beta \equiv P_{\beta+1} \equiv P_{\beta+2}, \equiv \dots\dots$$

We can thus state the following theorem:

If $P_1, P_2, \dots P_n, \dots P_\beta, \dots P_\alpha, \dots$ are all closed sets such that any one contains all the others with a higher index; then, either P_β vanishes for some definite number β of the first, or the second, class, or else there is a definite number β, from and after which all the sets are identical.

83. *Every perfect linear set* has the cardinal number c of the continuum; and every closed infinite set has the cardinal number c, or else the cardinal number a of the rational numbers.*

Let the intervals whose internal points are the set $C(G)$, the complement of the perfect set G, be denoted by $\{\delta\}$; and let Δ denote the greatest, or one of the greatest in case of equality, of the intervals $\{\delta\}$. Let l, the whole interval (a, b) in which G lies, be divided into the three parts l_0, Δ, l_1 so that $l = l_0 + \Delta + l_1$, where l_0 is on the left, and l_1 on the right of Δ, the greatest interval of $\{\delta\}$. Denote the greatest of the intervals $\{\delta\}$ in l_0, by Δ_0, and the greatest in l_1, by Δ_1; then the interval l_0 is divided by means of Δ_0 into three parts l_{00}, Δ_0, l_{01} in order from left to right, and the interval l_1 is divided by means of Δ_1 similarly into l_{10}, Δ_1, l_{11}. Proceeding in this manner to a further subdivision, let Δ_{pq} be the greatest of the intervals $\{\delta\}$ which lie in l_{pq}, where p, q each has one of the values 0 or 1; then l_{pq} is divided into three parts $l_{pq0}, \Delta_{pq}, l_{pq1}$; and so on indefinitely. The intervals $\{\delta\}$ are thus arranged in the order

$$\Delta, \Delta_0, \Delta_1, \Delta_{00}, \Delta_{01}, \Delta_{10}, \Delta_{11}, \dots$$

---

* Cantor, *Math. Annalen*, vol. XXIII (1884), pp. 486–488.

and each interval of $\{\delta\}$ occurs at a definite place in the sequence. Consider a sequence of intervals $l, l_p, l_{pq}, l_{pqr}, \ldots$, where $p, q, r, \ldots$ all have definite values, each ot which is either 0 or 1. Each of these intervals is contained in the preceding one, and has one end-point in common with it; and the sequence determines a single point P which is interior to all the intervals of the sequence, unless, from and after some fixed index, all the indices are identical, in which case P is a common end point of all the intervals after a fixed one. Hence, since the point P is not interior to any of the intervals $\{\delta\}$, it is a point of G. Conversely, every point of G can be so determined by means of a sequence of intervals; for every point of G belongs either to l_0 or to l_1, and also to one of the four intervals $l_{00}, l_{01}, l_{10}, l_{11}$, and so on. The point P is the limiting point of the end-points of the intervals $\Delta_p, \Delta_{pq}, \Delta_{pqr}, \ldots$ with the indices the same as those of the sequence $l_p, l_{pq}, l_{pqr}, \ldots$ which determines the point.

Every number of the continuum (0, 1) is expressible in the dyad scale by means of a sequence $\cdot p, \cdot pq, \cdot pqr, \ldots$, where each of the numbers $p, q, r, \ldots$ is either 0 or 1; and all numbers are expressed uniquely in this manner, except those for which all the digits after some fixed one are 1, these numbers being also expressible by a sequence in which only 0 occurs after some fixed place. The numbers last mentioned correspond as indices of $l_p, l_{pq}, l_{pqr}, \ldots$ to a point of G which is an end-point of one of the intervals $\{\delta\}$; but in every other case a number in the dyad scale corresponds to a point of G which is not an end-point of the intervals $\{\delta\}$. Since the set of numbers of the continuum (0, 1) has the cardinal number c, it follows that the points of G form a set of the same cardinal number, because each point of G corresponds uniquely to a single number of the continuum, except that two points of G which are end-points of one contiguous interval correspond to a single number of the continuum. Every closed set which is not enumerable has been shewn to contain a perfect set as component; such a set has therefore the cardinal number c.

It will appear from the theory of order-types which will be discussed in Chapter IV that the set of intervals $\{\delta\}$ which define a perfect set G, when taken in their order of position from left to right, have an order-type which is the same as η the order-type of the rational numbers which lie between 0 and 1, excluding 0 and 1 themselves, taken in their natural order in the continuum. It follows that a correspondence can be established between the intervals and the rational numbers, in which any two intervals correspond to two rational numbers that have the same order. If we take each rational number to correspond to the end-points of the corresponding interval, then each irrational number corresponds to a point of G which is a limiting point of end-points of intervals.

EXAMPLES

1. Let x be a number given by $x=\frac{c_1}{3}+\frac{c_2}{3^2}+\frac{c_3}{3^3}+\dots+\frac{c_n}{3^n}+\dots$, where the numbers $c_1, c_2, \dots c_n, \dots$ have each one of the values 0, 2. The set $\{x\}$ is a non-dense perfect set.

No number of the set lies between

$$\frac{c_1}{3}+\frac{c_2}{3^2}+\dots+\frac{0}{3^n}+\frac{2}{3^{n+1}}+\frac{2}{3^{n+2}}+\dots \text{ or } \frac{c_1}{3}+\frac{c_2}{3^2}+\dots+\frac{1}{3^n},$$

and

$$\frac{c_1}{3}+\frac{c_2}{3^2}+\dots+\frac{2}{3^n};$$

these two numbers determine a complementary interval of the set, the interval being of length $\frac{1}{3^n}$. The number of complementary intervals of length $\frac{1}{3^n}$ is 2^{n-1}, hence the sum of all the complementary intervals is $\sum_{n=1}^{\infty}\frac{2^{n-1}}{3^n}$, which is unity. It is clear that the set of complementary intervals is everywhere dense, and thus the set of points is non-dense. This example was constructed* by Cantor, and is the first example of a perfect non-dense set which has been purposely constructed.

2. Let us suppose that the numbers of the interval (0, 1) are expressed in the dyad scale, in the form $a_1a_2a_3 \dots a_n \dots$; where each a is either 0 or 1. Each number for which the a's all vanish, after some fixed one a_n, which must be 1, is also representable as an unending radix-fraction, in which a_n is 0, and all the subsequent digits are 1. Let the numbers now be interpreted as if they were in the decimal scale. To each irrational number in the dyad scale, there corresponds a single number in the decimal scale, represented by the same digits. Of each rational number, not represented by a recurring radix-fraction, there is a double representation in the dyad scale, and there correspond two numbers in the decimal scale, which define a complementary interval of the set of points which represents the numbers in the decimal scale. A perfect non-dense set of points is thus defined.

3. Taking a positive integer m (>2), let the interval (0, 1) be divided into m equal parts, and exempt the last part from further subdivision. Divide each of the remaining $m-1$ intervals into m equal parts, and in each case exempt the last part from further subdivision. Let this operation be continued indefinitely. The points of division form a non-dense set; for if an interval d be taken anywhere in the interval (0, 1), k may be so chosen that $\frac{1}{m^k}<\frac{d}{2}$, and a segment $\left(\frac{a}{m^k}, \frac{a+1}{m^k}\right)$ entirely within d, can be determined. This segment is either an exempted interval, or its mth part is one. The end-points of the intervals, together with their limiting points, form a non-dense closed set†, of cardinal number c.

4. As in† Ex. 3, let the interval (0, 1) be divided into m equal parts, and the last be exempted from further division. Then let the remaining $m-1$ parts each be divided into m^2 equal parts, the last of each being exempted from further division. Let the remaining parts be then divided into m^3 equal parts, the last of these in each case being exempted from further division. If this process be carried on indefinitely, the end-points of the divisions, together with their limiting points, form a non-dense closed set, of cardinal number c.

---

* See *Math. Annalen*, vol. XXI (1883), p. 590.

† See H. J. S. Smith, *Proc. Lond. Math. Soc.* (1), vol. VI (1875), pp. 147, 148.

5. Let $k_1, k_2, \dots k_n, \dots$ be a sequence of positive integers, each of which is greater than unity, and defined according to any law.

It can be shewn[*] that every irrational number x, in $(0, 1)$, can be uniquely represented in the form

$$x = \frac{c_1}{k_1} + \frac{c_2}{k_1 k_2} + \dots + \frac{c_n}{k_1 k_2 \dots k_n} + \dots,$$

where $c_n < k_n$, and not all of the numbers $c_n, c_{n+1}, \dots$ are zero, for any value of n.

It can further be shewn that

$$x = 1 - \frac{\eta_1}{k_1} - \frac{\eta_2}{k_1 k_2} - \dots - \frac{\eta_n}{k_1 k_2 \dots k_n} - \dots,$$

where $\eta_n = k_n - 1 - c_n$. If, from and after a certain value of n, the condition $c_n = k_n - 1$ is always satisfied, then all the η_n vanish, and x is rational. It thus appears that these rational numbers are capable of a double representation in the form

$$x = \frac{c_1}{k_1} + \frac{c_2}{k_1 k_2} + \dots + \frac{c_n}{k_1 k_2 \dots k_n} + \dots;$$

(1) by the vanishing of all the c, after some fixed one, and (2) by the condition $c_n = k_n - 1$ being satisfied, from and after some fixed value of n.

If we now take those values of x, for which every c does not exceed some fixed integer λ, these values of x form a non-dense perfect set G_λ. It is easily seen that the interval of which the end-points are

$$\frac{c_1}{k_1} + \frac{c_2}{k_1 k_2} + \dots + \frac{c_n}{k_1 k_2 \dots k_n} + \frac{\lambda}{k_1 k_2 \dots k_{n+1}} + \frac{\lambda}{k_1 k_2 \dots k_{n+2}} + \dots$$

and

$$\frac{c_1}{k_1} + \frac{c_2}{k_1 k_2} + \dots + \frac{c_n + 1}{k_1 k_2 \dots k_n}$$

contains no points of the set in its interior, although these points belong to the set.

A particular case of this set consists of the numbers given by

$$x = \frac{c_1}{10} + \frac{c_2}{10^{1 \cdot 2}} + \frac{c_3}{10^{1 \cdot 2 \cdot 3}} + \dots + \frac{c_n}{10^{n!}} + \dots,$$

where every c is ≤ 9. This set consists of the transcendental numbers first defined by Liouville[†].

PROPERTIES OF THE DERIVATIVES OF LINEAR SETS

84. If a set is dense in any sub-interval of the interval in which it is contained, its derivative $G^{(1)}$ contains every point of the sub-interval, and is identical, so far as such sub-interval is concerned, with the totality of the points of the sub-interval; we confine ourselves therefore to the case in which G is a non-dense set, and consequently its derivatives are also non-dense.

The derivatives of transfinite orders have been defined in § 68; and it was there shewn that there is either a last derivative, whose order is some number of the first class, or non-limiting number of the second class; or else that derivatives of all such orders exist, and have a set of points $G^{(\Omega)}$ in common.

---

[*] Brodén, *Math. Ann.* vol. LI (1899), p. 299.

[†] *Liouville's Journal*, vol. XVI (1851), p. 133.

It was shewn in § 68, that $G^{(1)}$ being a non-dense closed set, two cases arise:

(1) If $G^{(1)}$ is enumerable, in which case G is also enumerable, then $G^{(\beta)}$ vanishes for some number β of the first or the second class. A set with an enumerable derivative is called a *reducible* set.

(2) If $G^{(1)}$ is not enumerable, then there exists some number β, of the first or second class, for which $G^{(\beta)}$ is a perfect set, and is consequently identical with $G^{(\beta+1)}$, and with $G^{(\Omega)}$ as defined in § 68. The set $G^{(1)}$ is the sum of an enumerable set and the perfect set $G^{(\beta)}$. A set G which has this property, is said to be *irreducible*.

It should be observed that, when $G^{(1)}$ is unenumerable, and consequently of cardinal number c, the same as the cardinal number of its perfect component, we are unable to make any inference as to the cardinal number of G itself. This may be a or c, or other cardinal number between the two, in case such a number exists.

CLOSED SETS IN TWO OR MORE DIMENSIONS

85. In considering the properties of closed sets of points in space of two or more dimensions it is sufficient to treat in detail the case of plane sets only, because these sets exhibit sufficiently clearly the respects in which linear closed sets differ from other closed sets. The results obtained for plane closed sets can immediately be extended to the case of such sets in any number of dimensions. In the case of a linear set, each point P which does not belong to a given closed set is enclosed in an open interval which contains no points of the set, and this interval has a maximum length in both directions, the end-points of such maximum interval δ being points of the closed set (or an end-point of the fundamental interval), and this maximum interval is identical with δ, for all points interior to δ. But, in the case of a plane set, if we confine ourselves to areas of given form and orientation, such as rectangular cells, and these take the place of the linear intervals δ, it is not the case that a closed set is defined as the set of boundary points, together with the external points, of a unique system of such cells.

If P be a point which does not belong to a given plane closed set G, in a fundamental cell, and if we draw through P a straight line parallel to the line whose equation is $y = mx$, then those points of the given set which lie on this straight line form a closed set (see § 56), and the point P must be interior to an interval $\Delta_m(P)$ contiguous to this closed set. If, on either side of P, there are in this straight line no points of G, then on this side the extremity of the interval $\Delta_m(P)$ may be regarded as the point in which the straight line intersects a side of the fundamental cell. The interval $\Delta_m(P)$ exists for every value of m, and the extremities of the interval, for a fixed P, are points of the plane set, or points on the boundary of the fundamental cell. The region of plane space Δ_m, defined by all these intervals, for every value of m, when that straight line is included for which x has a constant value, is free in its interior from points of G. Such a region $\Delta_m (P)$ may be regarded as the analogue, for plane sets, of the interval $\delta (P)$ for the case of linear sets. But the analogy is not complete, for if Q be an interior point of $\Delta_m (P)$, it is not necessarily the case that $\Delta_m (Q)$ is identical with $\Delta_m (P)$.

If however we work only with rectangular cells, there exists in general no unique rectangular cell, corresponding to a point P of $C' (G)$, which is such that no interior point of it is a point of G, and that on its boundary there is a point of G. If we describe a square cell of sides 2ρ with its centre at P, and containing in its interior and on its boundary no point of G, we may keep three of the sides fixed, and move the fourth parallel to itself until it either contains a point of P, or becomes coincident with one of the sides of the fundamental rectangle. We may then proceed to move another side parallel to itself until the same thing happens, and so on with the other two sides. The resulting rectangle will in general depend upon the order in which the moving of the four sides takes place. Thus we obtain, in general, no unique rectangular cell corresponding to the point P.

86. It is however possible, for a given closed non-dense plane set G, to construct an enumerable set, not unique, of cells which is everywhere dense, and such that every point of G lies on the boundary of a rectangle, or is a limiting point of points which lie on the boundaries of such cells Let us denote by S the fundamental cell in which the whole set G lies, and let δ be a cell constructed, as above, for a point P of the set $C' (G)$. Produce the sides of δ, when necessary, until they cut the sides of S, thus dividing S into at most nine different rectangles, of which one is δ, and the others may be denoted by S_r, where $r = 1, 2, \dots 8$. In each rectangle S_r take any point P_r which does not belong to G, and construct for P_r a free rectangle δ_r, as before; let the sides of δ_r be produced, when necessary, until they meet the sides of S_r, then S_r is divided into at most nine rectangles, which consist of δ_r and at most eight rectangles S_{rs}, where $s = 1, 2, \dots 8$.

Proceeding in this manner, we obtain a set of rectangles

$$S, S_r, S_{rs}, S_{rst}, \dots,$$

and in them a set of rectangles $\delta, \delta_r, \delta_{rs}, \delta_{rst}, \dots$, each of which contains no points of G in its interior; each of the numbers $r, s, t, \dots$ being one of the digits $1, 2, 3, \dots 8$. If p be a point of G which is not on a boundary of any rectangle δ_n, it must be in the interior of each of an unending sequence of rectangles $S_r, S_{rs}, S_{rst}, \dots$, where $r, s, t, \dots$ have definite values; and this set of rectangles must converge either (1), to a point in the interior of all of them, or (2), to a linear interval, or (3), to a definite rectangle S_ω in the interior of all of them. In case (1), the point to which the rectangles converge is a limiting point of those points of G which lie on the boundaries of the definite sequence of rectangles δ_r, δ_{rs}, δ_{rst}, ... In case (2), there must be, in the limiting linear interval, at least one point which is a limiting point of G: for, if not, the whole interval could be enclosed in a rectangle which contains no points of G; and this is impossible. In case (3), we start with the rectangle S_ω, and take a point P_ω not belonging to G inside it, which point exists, since G is non-dense, and construct the maximum free rectangle δ_ω. Produce its sides as before to meet those of S_ω, and proceed as before to construct $S_{\omega rst}$ ... and $\delta_{\omega rst}$ .... This process can be continued until an index is reached which may be any number of the second class, but the point p must be reached before some definite number of the second class appears as index; this following from the fact that the number of non-overlapping regions which are contained in a given space must be enumerable. Thus the point p is reached after an enumerable set of steps of the process.

It has therefore been shewn that

If G is a non-dense closed plane set of points, an everywhere dense enumerable set of rectangles exists, such that every point of G is on a boundary of one or more of the rectangles, or is a limiting point of such points, or lies in a linear interval which is the limit of a sequence of the rectangles.

In case the set G is perfect, the rectangles of the set must either not abut on one another, or every common side must contain either no points of G, or else a perfect set of points of G.

87. That a perfect plane set G has the power of the continuum* may be proved by projecting the points of G on a side of the fundamental rectangular cell. The linear set of points which are the projections of G is a closed set. For if P be a limiting point of the set of projects, let pp' be an arbitrarily small neighbourhood of P (on the side), of which P is the centre; construct straight lines PQ, pq, $p'q'$, perpendicular to pp', meeting the opposite side of the fundamental rectangle in Q, q, q'. In the rectangle $pqq'p'$ there are an infinite number of points of G. Let a symmetrical system of nets, with closed meshes, be fitted on to this rectangle $pqq'p'$; the number m (see § 51), whose square is the number of meshes in D_1, being taken to be odd. There is in at least one mesh of D_1 an infinite number of points of G in the interior or on one of its boundaries parallel to pp'; and one such mesh at least must exist with its centre on PQ; for otherwise P could not be a limiting point of the projection of G. We take that one d_1 of such meshes which is nearest to pp'. Similarly there exists a mesh d_2, of D_2, contained in d_1, with its centre on PQ, and containing an infinite set of points of G. Proceeding in this manner, we obtain a sequence $d_1, d_2, \ldots$ of meshes each containing the next, and all having the same property as d_1. These define a point on PQ which must be a limiting point of G, and therefore belongs to G. Therefore P is a point of the projection of G; and thus the projection of G is a closed set. If a point P be an isolated point of the projection of G, we may as before construct a rectangle $pqq'p'$ which contains no points of G that do not lie on PQ. The component of G in this rectangle must be perfect, and therefore the straight line PQ contains a perfect component of G. If the projected set is perfect, then it has the power c of the continuum; and if it contains isolated points, these must be the projections of perfect linear components of G; therefore, in either case, G has the power of the continuum.

---

* See Bendixson, *Bib. Svensk. Vet. Handl.* vol. IX (1884), where the first proof of this theorem was given

It is clear that this method of proof can be extended to the case of a set in any number of dimensions.

THE ANALYSIS OF SETS IN GENERAL

88. With a view to the general analysis of sets of points in any number of dimensions, it is necessary to classify the points of a given set according to the cardinal number of those points of the set that are in the arbitrarily small neighbourhoods of the various points[*].

An isolated point of any set G is such that in a sufficiently small neighbourhood of the point there are no other points of G. For this reason an isolated point may be said to be of *degree zero in the set*.

A limiting point of G, such that in every sufficiently small neighbourhood of the point there is an enumerably infinite set of points of G, is said to be a point of *enumerable degree in the set*, or of *degree a in the set*. Such a point may or may not belong to G.

In case the point is such that, in every neighbourhood of it, there exists a set of points of G that has the cardinal number of the continuum, the point is said to have degree c in the set.

As we are not entitled to assert that every infinite set of points has either a or c for its cardinal number, we contemplate the existence of sets having a hypothetical cardinal number x different from a or c. A point such that, in every sufficiently small neighbourhood of the point, there existed a set, of cardinal number x, of points belonging to G, would be termed a point of *degree x in the set G*.

---

[*] An analysis of this kind, for sets in general, was given by Cantor, *Acta Math.* vol. VII (1885), p. 105. A more elementary presentation of the matter, without the use of transfinite numbers, has been given by W. H. Young, *Quart. Journ. of Math.* vol. XXXV (1904), p. 102.

A point P, whether it belongs to G or not, is said to be a point of *unenumerable degree in* G if, in every neighbourhood of the point, there exists an unenumerable set of points which belong to G. Such a point has been termed* by Lindelöf a *point of condensation*. As however this expression is employed by some writers to denote a limiting point, it is better avoided. The set of all such points P (belonging to G or not) is closed, for any limiting point clearly belongs to the set. It is obvious that, if G be enumerable, it contains no points of unenumerable degree in G, for every part of an enumerable set is enumerable (finite or infinite).

Conversely it may be shewn that:

If no point of G is a point of unenumerability in G, the set G is enumerable.

Let a system of nets be fitted on to the finite, or indefinitely great, cell in which G is contained. Any point P, of G, is defined by a unique sequence of meshes of the successive nets of the system. There must be a value n_P of n, from and after which all the meshes $\{d_n\}$ of the sequence which defines P contain only an enumerable set of points of G. Thus, to each point P of G, we can correlate one mesh d_{n_P}; but the same mesh may be correlated with more than one point P. Taking all the meshes d_{n_P}, for all the points P of G, these form an enumerable set, being part of the enumerable set of all the meshes of the system of nets. Since each mesh d_{n_P} contains only an enumerable set of points of G, it follows (§ 58) that G is an enumerable set.

89. *Every set G, not enumerable, is the sum of an enumerable set (possibly absent) and a set in which each point is of unenumerable degree in the set, and which is therefore dense in itself.*

Thus $G = H + K$, where H is enumerable, and K is such that each point is a point of unenumerable degree in G. Since every point of K has in its neighbourhood an unenumerable set of points of G, and therefore of K, the set K is dense in itself. To prove the theorem, let H be that part of G, each point of which is a point of zero, or of enumerable, degree in G; it is then clear that H must be a set in which every point is of zero, or of enumerable, degree in H; therefore H is enumerable. The set $K = G - H$ is then unenumerable because each point of it contains in its neighbourhood an unenumerable set of points of G, and therefore of K.

It is clear that every limiting point of a set of points of unenumerability is itself such a point. Since K is dense in itself, it is contained in its derivative K'; and as K' has no isolated points, it is perfect.

If G is a closed set, it must contain all the points which are of unenumerable degree in it; thus K' and K are identical, and therefore K is perfect. Therefore we have the theorem:

---

* See *Comptes Rendus*, vol. CXXXVII (1903), p. 697.

A closed set is the sum of an enumerable set and a perfect set, either of which may be absent.

The proof of the theorem here given is due to Lindelöf (*loc. cit.*).

For the special case of linear closed sets this theorem has already been proved in § 81. But the following theorem may here be added[*]:

A linear closed set is the sum of an enumerable set and a set which is dense in itself, such that each point of the latter set is a limiting point on both sides.

For, a point of K that is a limiting point only on one side is also a point of K' which can only be a limiting point of K' on one side; moreover the set of all such points of K', viz. the extremities of the intervals contiguous to it, is enumerable. Therefore those points of K that are limiting points of K on one side only form an enumerable set; and if these be removed from K we have a set K_1 the points of which are limiting points of K, and therefore also of K_1, on both sides.

It is asserted in the first theorem given above that every unenumerable set contains a component that is dense in itself. An enumerable set may contain such a component, and if it does, its derivative contains a perfect set, and thus the enumerable set is irreducible.

It has been proved by Sierpiński[†] that every set of points can be decomposed into the sum of two sets, the first of which contains no component that is dense in itself, and is effectively enumerable, and the second of which is dense in itself. Either set may be absent.

90. *If G is an unenumerable set, those points of G which are of the same degree x ($> a$) in G form a set which is dense in itself, and of which the cardinal number is $\geq x$.*

Let us fit on a system of nets with closed meshes to the unbounded space in which G is contained. Consider any point P that is of degree x in G; P is in a mesh $d_n(x)$ in which the set of points of G is of degree x; n may have the smallest value for which this is the case. When all such points P are taken, the meshes $d_n(x)$ corresponding to them form an enumerable (or finite) set, a part of the enumerable set of all the meshes of the system of nets. In each of these meshes $d_n(x)$ there is a set of points of G of cardinal number x. It will be shown in Chapter IV, § 155, in connection with the general theory of cardinal numbers, that an enumerable set of aggregates of points, each of cardinal number x, cannot be assumed to have the cardinal number x, unless $x = c$. Since all those points of G that are of degree x in the set are contained in the enumerable set of meshes so obtained, it follows that the set of such points of G has a cardinal number $\geqq x$, except that it certainly has the cardinal number c, in case $x = c$.

---

[*] W. H. Young, *Quart. Journ. of Math.* vol. xxxix (1908), p. 76.

[†] *Fundamenta Mat.* vol. I (1920), p. 1.

To shew that the set is dense in itself, we observe that, in a sufficiently small neighbourhood (α, β) of P_x, there is a set of points of G, of cardinal number x; and that none of these points can be of degree in G higher than x. For if there were such a point Q of degree higher than x, in some interval (α', β') contained in (α, β) and containing Q in its interior, there would be a set of points of G of cardinal number higher than x; but this is impossible, as all these points would be in (α, β). Again, the points of G in (α, β), other than P_x, cannot be all of degree lower than x; for if they were so, their cardinal number would be lower than x, since they could be enclosed in an enumerable set of non-overlapping intervals with P_x as sole external point. Moreover, since, in any arbitrarily small neighbourhood of P_x, there are points of the same degree x in G, P_x is a limiting point for such points. Therefore the set of points such as P_x is dense in itself.

The set of points of degree c is dense in itself, and of cardinal number c.

91. Any set G consists of isolated points which form an enumerable set called the *adherence* of G, and of limiting points which form a set called the *coherence* of G. Denoting the adherence and the coherence of G by Ga, Gc respectively, we have $G = Ga + Gc$.

The set Gc can in a similar manner be split up into its adherence and its coherence, which we denote by Gca and Gc^2 respectively, thus

$$Gc = Gca + Gc^2.$$

The set Gca is an isolated set, and therefore enumerable; and if we proceed to resolve Gc^2 in a similar manner into its adherence Gc^2a, and its coherence Gc^3, and then to resolve Gc^3, it is clear that the process may be continued any number n of times. We thus obtain

$$G = Ga + Gca + Gc^2a + \ldots + Gc^{n-1}a + Gc^n.$$

The set $Gc^{n-1}a$ may be named the adherence of G of order n, and Gc^n may be denominated the coherence of G of order n.

It may happen that, for some value of n, Gc^n vanishes; in that case G has been split up into a finite number of enumerable sets, and is consequently itself enumerable. If this be not the case, the process may be continued indefinitely, and Gc^n then exists for every value of n. We then define

$$D\,(G, Gc, Gc^2, \ldots Gc^n, \ldots),$$

the set of points common to all the coherences of G, to be the coherence of order ω, and denote it by Gc^{ω}. It is clear that every point of G which does not belong to one of the sets $Gc^{n-1}a$, belongs to Gc^{ω}, hence we have

$$G = \Sigma Gc^{n-1}a + Gc^{\omega},$$

the summation being taken for all values of n belonging to the first class.

We now split up Gc^{ω} into its adherence $Gc^{\omega}a$, and its coherence $Gc^{\omega+1}$, and proceed further to obtain the adherences and coherences of G of the orders of the various numbers of the second class. If $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$ is a sequence of numbers of the second class, which has β for its limit, the coherence of order β is defined by

$$Gc^{\beta} = D\,(Gc^{\alpha_1}, Gc^{\alpha_2}, \ldots Gc^{\alpha_n}, \ldots).$$

We now obtain a resolution of G of the form

$$G = \Sigma Gc^{p}a + Gc^{\gamma},$$

where γ is any number of the first or second class, and the summation refers to all values of p which are less than γ. Each adherence $Gc^{p}a$ is an isolated set, and therefore enumerable, and if G contains a component which is dense in itself, this component is contained in Gc^{γ}.

First suppose G to be an enumerable set; the process of analysis must then cease for some number γ of the first or second class. For if $Gc^{\gamma}a$ existed for every number γ of the second class, we should have obtained an unenumerable set of adherences containing no points in common, and all belonging to G: thus G could not be enumerable.

The cessation of the process may take place in two different manners:

(1), if for some number γ of the first or second class, $Gc^{\gamma} = 0$, G has been resolved into an enumerable set of adherences, and it contains no component which is dense in itself:

(2), if for some number γ,

$$Gc^{\gamma}a = 0,$$

in which case $Gc^{\gamma} = Gc^{\gamma+1}$, the set Gc^{γ} then contains no adherence, and every point of it is a limiting point, and Gc^{γ} is therefore dense in itself. The set G has consequently been resolved into an enumerable component which contains no part that is dense in itself, and into a set which is enumerable and dense in itself.

Next, let us suppose that G is an unenumerable set. Then it has been shewn that those points of G which are of unenumerable degree in G form a set that is dense in itself; and those points which belong to the adherences of all orders are points of zero, or of enumerable, degree, and thus form an enumerable set. It follows, since all points that do not belong to that part of G which is dense in itself belong to the adherences, that the number of adherences must be enumerable; and thus that, for some number γ of the first or second class, Gc^{γ} is dense in itself. The set

Gc^{γ} may consist of an enumerable set dense in itself, and of sets of higher cardinal numbers dense in themselves.

It has thus been shewn that *any set G may be represented by*

$$G = U + V_a + \Sigma V_x + V_c,$$

where U is an enumerable set which contains no component that is dense in itself, V_a is an enumerable set of points of degree a dense in itself, V_c is a set, of cardinal number c, consisting of points of degree c, dense in itself; V_x is a set dense in itself consisting of points of degree x, where $a < x < c$.

If, as is probable, no cardinal numbers exist between a and c, the sets V_x can be omitted. A set such as V_a, or V_x, or V_c is denominated a homogeneous set of degree a, or x, or c, in the set G.

If G is a closed set, then, as has been shewn in § 89, V_c is perfect, and ΣV_x cannot exist.

INNER AND OUTER LIMITING SETS

92. Let $S_1, S_2, \ldots S_n, \ldots$ be a sequence of sets of points, in one or more dimensional space, such that each set S_{n+1} of the sequence is contained in the preceding one S_n, then the set S_ω, or $D(S_1, S_2, \ldots S_n, \ldots)$, consisting of points each of which belongs to all the sets of the sequence, is said, when such set exists, to be the *inner limiting set* of the sequence of sets. A point which does not belong to $S_n, S_{n+1}, \ldots$ but belongs to S_{n-1} is said to be shed at the index n.

When $S_1, S_2, \ldots S_n, \ldots$ is a sequence of sets, each one of which S_n is contained in the next S_{n+1}, the set S_ω, or $M(S_1, S_2, \ldots S_n, \ldots)$, which consists of the set of those points each of which belongs to all the sets, from and after some value of n dependent on the particular point, is said to be the *outer limiting set* of the sequence.

If any sequence $\Sigma_1, \Sigma_2, \ldots \Sigma_n, \ldots$ of sets be given, and we take $S_1 = \Sigma_1$, $S_2 = D(\Sigma_1, \Sigma_2)$, $S_3 = D(\Sigma_1, \Sigma_2, \Sigma_3)$, and in general $S_m = D(\Sigma_1, \Sigma_2, \ldots \Sigma_m)$, the inner limiting set S_ω, when it exists, also defines $D(\Sigma_1, \Sigma_2, \ldots \Sigma_m, \ldots)$, and may be regarded as defined by the given sets $\Sigma_1, \Sigma_2, \ldots$.

For a point that belongs to all the sets $\Sigma_1, \Sigma_2, \ldots \Sigma_n, \ldots$ belongs to all the sets $S_1, S_2, S_3, \ldots S_n, \ldots$; and a point that belongs to all the latter sets belongs to all the former sets.

Thus it has been shewn that *the set of points common to all the sets of any given sequence of sets is an inner limiting set.*

Again, if we take $S_1 = \Sigma_1$, $S_2 = M(\Sigma_1, \Sigma_2)$, $S_3 = M(\Sigma_1, \Sigma_2, \Sigma_3)$, … the outer limiting set S_ω also defines $M(\Sigma_1, \Sigma_2, \ldots \Sigma_n, \ldots)$, and thus *the set of all points that belong to one at least of the sets of any given sequence of sets is an outer limiting set.*

If S_ω be the inner limiting set of the sequence $S_1, S_2, S_3, \dots$ of sets, each of which contains the next, the sequence $C(S_1), C(S_2), C(S_3), \dots$ of the sets complementary to the first sequence with respect to a cell (or interval), whether finite or not, which contains them all, is such that each set is contained in the next, and their outer limiting set is $C(S_\omega)$.

It has been shewn in § 67 that, when the sets $S_1, S_2, \dots$ are all closed sets, the inner limiting set always exists, and is a closed set

Every set of points can be exhibited as the inner limiting set of a sequence of sets each of which contains the next; and also as the outer limiting set of a sequence of sets each of which is contained in the next.

Let a set G be contained in the interior of a cell $(a^{(1)}, a^{(2)}, \dots a^{(p)}, b^{(1)}, b^{(2)}, \dots b^{(p)})$; and let G_θ be the component of G that is contained in the cell $(a^{(1)}, a^{(2)}, \dots a^{(p)}; \theta b^{(1)}, \theta b^{(2)}, \dots \theta b^{(p)})$, where $0 < \theta < 1$.

If $\theta_1, \theta_2, \dots \theta_n, \dots$ be an increasing sequence of values of θ that converges to 1, the sets $G_{\theta_1}, G_{\theta_2}, \dots$ will be such that each one is contained in the next, and their outer limiting set is the given set G.

Also G is the inner limiting set of the sequence of those sets which are complementary to the sets of a sequence for which $C(G)$ is the outer limiting set.

In case G is unbounded, it may be placed in correspondence with a set interior to the cell $(-1, 1)$, and the theorem therefore holds for G, since it holds for the set that corresponds to G.

SETS OF THE FIRST, AND OF THE SECOND, CATEGORY

93. *The outer limiting set of a sequence of non-dense sets $G_1, G_2, \dots G_n, \dots$ each of which contains the preceding one is said to be a set of the first category.*

This is equivalent to the definition given* by Baire, that, if $P_1, P_2, \dots P_n, \dots$ be any sequence of sets, non-dense in the fundamental interval, or cell, in which they are all contained, the set $M(P_1, P_2, \dots P_n, \dots)$, consisting of all the points that belong to any of the given sets, is of the first category. For we have only to take $G_n = M(P_1, P_2, \dots P_n)$; and G_n is non-dense. The set $P_1, P_2, \dots P_n, \dots$ can always be replaced by non-dense sets, $Q_1, Q_2, Q_3, \dots$ no two of which have a point in common; for we may take

$$Q_1 = P_1,\ Q_2 = P_2 - D(P_1, P_2),\ Q_3 = P_3 - D(Q_1, P_3) - D(Q_2, P_3), \dots.$$

Any set which is not a set of the first category is said to be of the second category.

---

* *Annali di Mat.* (3), vol. III (1899), p. 65, where the distinction between sets of the first and second categories was first introduced.

It is clear that a set of the first category is enumerable if the sets $G_1, G_2, \ldots G_n, \ldots$ are all enumerable, and that every enumerable set is of the first category.

An enumerable set of the first category may be everywhere dense. For example, the set of rational points in a given cell, or interval, (a, b) is of the first category, and is everywhere dense.

A set of the first category may have the cardinal number of the continuum, and may then be everywhere dense, or non-dense.

A set of the first category is not necessarily closed. As an illustration of this fact, let us consider a set of cells $P_1Q_1, P_2Q_2, \ldots P_nQ_n, \ldots$ each of which contains the next, and such that the sequence converges to a single point p interior to all the cells of the sequence. Let G_1 consist of a set of points on the boundary of P_1Q_1; let G_2 consist of G_1 together with a set of points on the boundary of P_2Q_2; and generally, let G_n consist of G_{n-1} together with a set of points on the boundary of P_nQ_n. The point p does not belong to G_n for any value of n, and therefore does not belong to G_ω, but it is clearly a limiting point of G_ω; and therefore G_ω is not a closed set.

To illustrate the fact that an unenumerable set of the first category may be everywhere dense, fit on to the cell, or interval, (a, b) a system of closed nets. Let G_1 be a non-dense perfect set in (a, b); in each mesh of D_1 place a non-dense perfect set, and let G_2 be the perfect set made up of all such non-dense perfect sets, together with G_1. Let G_3 consist of G_2 together with non-dense perfect sets placed in all the meshes of D_2; and so on.

The set of the first category defined by $G_1, G_2, \ldots$ is clearly everywhere dense; for any sub-cell contains a mesh of D_n, for some sufficiently large value of n, and thus contains points of G_n, and therefore of G_ω.

In case G_ω is closed, it must be non-dense, for it will be shewn that $C(G_\omega)$ is everywhere dense, and therefore G_ω cannot contain a cell, or interval.

94. *A set of points which is complementary to a set of the first category is a set of the second category.*

In particular, a cell, or an interval, is not the sum of two sets of the first category.

To prove this theorem, we shew, in the first place, that the set complementary to a set of the first category is everywhere dense. For let (α, β) be any cell, or interval, in the fundamental cell, or interval, in which the set of the first category is contained. There exists a cell, or interval, (α_1, β_1) interior to (α, β), in which there is no point of the non-dense set G_1. Again (α_1, β_1) contains in its interior a cell, or interval, (α_2, β_2) containing no points of G_2; and so on.

There exists then a point interior to all the cells, or intervals,

$$(\alpha, \beta)\ (\alpha_1, \beta_1)\ (\alpha_2, \beta_2) \ldots\ldots (\alpha_n, \beta_n) \ldots\ldots$$

which is not a point of G_ω. Hence the set complementary to G_ω is everywhere dense. It follows from this that a set of the first category cannot contain all the points of any interval, or cell, and is thus a diffuse set. Next, let us assume, if possible, that $C(G_\omega)$ is itself the limit of a sequence $Q_1, Q_2, \ldots Q_n, \ldots$ of non-dense sets, each of which is contained in the next. The sets $G_1 + Q_1, G_2 + Q_2, \ldots$ are all non-dense, and each one of them is contained in the next; their limit, which is of the first category, is identical with the fundamental cell, and this has been shewn to be impossible. Hence the complement of a set of the first category is not of the first category.

The converse of the above theorem does not hold good. It is not true that every set of the second category is the complement of a set of the first category. Thus there are two kinds of sets of the second category; those which are complementary to a set of the first category, and those for which this is not the case. If, for example, in a linear interval (a, b) the points of a sub-interval (α_1, β_1) be taken, these make up a set of the second category, and the complementary set is also of the second category. It has been shewn* by Mahlo that the points of a linear interval (a, b) can be divided into two sets, each of which is everywhere dense, and of the cardinal number of the continuum, and neither of which is of the first category.

95. A set of the second category which is complementary to a set of the first category is said† to be a *residual* set.

A residual set can be obtained by the successive removal from a fundamental cell, or interval, of non-dense sets of points belonging to a sequence of such sets. It has been shewn in § 94 that a residual set is everywhere dense. It will now be shewn that:

A residual set has the cardinal number of the continuum.

Let G_ω be a set of the first category, and $C(G_\omega)$ the corresponding residual set. Let a system of nets be fitted on to the fundamental cell, or interval, in which G_ω is contained. If n be sufficiently large, there is a mesh of the net D_n that contains no point of G_1; the smallest value of n, say n_1, for which this is the case may be chosen. Let this mesh be denoted by d. There is some smallest value of n, say n_2, such that d contains in its interior two (or more) meshes d_0, d_1, of the net D_{n_2}, neither of which contains a point of G_2. In case there are more than two such meshes, d_0, d_1 are those of lowest rank. Again, there is some smallest value of n, say n_3, such that both d_0 and d_1 contain within them two or more meshes of D_{n_3}, none of which contains a point of G_3. Let the two meshes in d_0 be denoted by d_{00}, d_{01}; and the two meshes in d_1 by d_{10}, d_{11}. Proceeding in this way, we can define a sequence of meshes $d_p, d_{pq}, d_{pqr}, \ldots$ each of which contains the next in its interior; where each of the indices $p, q, r, \ldots$ is either 0 or 1; and where the sequence $p, q, r, \ldots$ is defined in accordance with some set of rules The point interior to all the meshes of this sequence does not belong to G_ω, and is therefore a point of the residual. But such a sequence can be defined corresponding to each number of the continuum; leaving out those sequences in which all the indices, from and after a fixed one, are all 1, the correspondence is unique. Hence the residual must have the cardinal number c, of the continuum. Further, it can be proved that:

---

* See *Leipz Ber.* vol. LXV (1913), p. 283.

† See Denjoy, *Journal de Math.* (7), vol. I (1915), p. 123, who has introduced this terminology.

Any finite, or enumerably infinite, set of residuals have in common a set which is also a residual set.

Let $H^{(1)}, H^{(2)}, \ldots$ be residuals obtained by removing successively the sets of sequences $\{G_n^{(1)}\}, \{G_n^{(2)}\}, \ldots$ of non-dense sets from an interval, or a cell; the non-dense sets can be arranged in enumerable order

$$G_1^{(1)}, G_1^{(2)}, G_2^{(1)}, G_1^{(3)}, G_2^{(2)}, G_3^{(1)}, \ldots.$$

If these sets be removed successively from the interval, or cell, we obtain the residual

$$D(H^{(1)}, H^{(2)}, H^{(3)}, \ldots).$$

96. It is often of importance to consider the properties of sets that are contained in a given perfect set H, which may be non-dense in the continuum.

If $P_1, P_2, \ldots P_n, \ldots$ be a sequence of sets all of which are non-dense in the perfect set H, the set $M(P_1, P_2, \ldots)$ is said to be of the first category relatively to H. A set which is not of the first category relatively to H is said to be of the second category relatively to H. A set whose complement relatively to H is of the first category relatively to H is said to be a residual set relatively to H.

It can be shewn that:

A set of points which is the complement, relatively to the perfect set H, of a set of the first category, relatively to H, is of the second category relatively to H.

In particular H cannot be the sum of two sets both of the first category relatively to H.

This is proved in a manner precisely similar to that in which the theorem for the case in which H is a continuous cell, or interval, was proved in § 94. Thus it is first shewn that a residual relatively to H is dense in H, and then the proof of the theorem is completed as in § 94.

It follows that a set of the first category relatively to H is diffused in H.

It can be proved, as in § 95, that a residual with respect to H has the cardinal number of the continuum.

It has been shewn by Denjoy[*] that a residual relatively to a perfect set H itself contains a perfect set.

It will appear, from the theory of order-types developed in Chapter IV. that the points of a perfect linear set can be made to correspond uniquely with points of a continuous interval (a, b), in such a manner that the relative order of two points of the perfect set is the same as that of the corresponding points in (a, b); the end-points of an interval contiguous to the perfect set corresponding to one point of (a, b). To a closed set H, non-dense in the perfect set, there corresponds a closed set, non-dense in the continuum; and sets of the first, or the second, category relatively to the perfect set correspond respectively to sets of the first, or of the second, category in (a, b). It thus appears that the properties of sets of the first, and the second, category in the continuum can be immediately extended to sets of the first, and the second, category relatively to a perfect set.

This is a particular case of the general property of any perfect set: viz. that all descriptive properties of sets of points in the continuum correspond to identical properties of sets of points in a perfect set, even if the perfect set be itself non-dense in the continuum.

EXAMPLES

1. Let $P_1, P_2, P_3, \ldots P_n, \ldots$ be an enumerable set of points in an interval (a, b); the set may be everywhere dense in (a, b). The finite sets

$$(P_1), (P_1, P_2), (P_1, P_2, P_3), \ldots (P_1, P_2, \ldots P_n), \ldots$$

are each closed, and the given set is the limiting set, which is therefore of the first category. The remaining points of (a, b) form a set of the second category, which is a residual.

2. Denoting the points of the interval $(0, 1)$, as in Ex. 5, § 83, by

$$x = \frac{c_1}{k_1} + \frac{c_2}{k_1 k_2} + \ldots + \frac{c_n}{k_1 k_2 \ldots k_n} + \ldots,$$

where $c_n < k_n$; let the fixed integers $k_1, k_2, \ldots k_n, \ldots$ form a sequence which increases without limit. If[†] $a_1, a_2, \ldots a_n, \ldots$ is any sequence of positive integers which increase without limit, let G_n denote the set of those numbers x, which are such that the integers $c_1, c_2, \ldots c_n, \ldots$ are all $< a_n$. The sets $G_1, G_2, G_3, \ldots G_n, \ldots$ are a sequence of perfect sets, each one of which contains the preceding ones; the set G_ω is then a set of the first category.

3. The numbers of the continuum $(0, 1)$ may be divided into sets, of the first, and of the second category, in the following manner: All the numbers in $(0, 1)$ may be expressed as endless decimals; the finite decimals being therefore not used. Let[‡] the set H consist of all those numbers in which the digit 9 occurs only a finite number of times, and of those numbers also in which, from and after some place, all the figures are 9. The comple-

---

[*] *Journal de Math* (7), vol. I (1915), p. 232.

[†] Brodén, *Math. Annalen*, vol. LI (1899), p. 299.

[‡] See Schoenflies, *Bericht*, vol. I, p. 106.

mentary set K consists of all those numbers in which 9 occurs an infinite number of times, except those in which every figure is 9, from and after some place. The set H is the limit of a sequence of non-dense closed sets $H_1, H_2, \dots H_n, \dots$ each of which is of cardinal number c. For, let H_1 consist of the numbers of the form $\cdot abc \dots k999 \dots$ in which every figure is 9, after some fixed place, and in which none of the figures $a, b, c, \dots k$ is 9; together with those decimals in which no figure is 9. No number of the set H_1 can lie within the interval $(\cdot abc \dots k899 \dots, \cdot abc \dots k999 \dots)$ which is therefore a complementary interval of the set. The set H_n may be taken to consist of the numbers of the form $\cdot abc \dots hk999 \dots$, in which k is not 9, and not more than $n-1$ of the figures $a, b, c, \dots h$, are 9; together with those decimals in which 9 does not occur. That each of the sets H_n is of cardinal number c, follows from the fact that it contains all the decimals in which 9 does not occur; and these, if interpreted in the scale of 8, represent all the numbers of the continuum $(0, 1)$. The set H is everywhere dense, since it contains that everywhere dense set of numbers in which every figure is 9, after some place. The set K, being of the second category, is also everywhere dense, and of cardinal number c.

4. The following method of dividing the continuum $(0, 1)$ into two portions, each of which is everywhere dense, and of cardinal number c, has been given by Brodén*: Let $l_0 + l_1 + l_2 + \dots + l_n + \dots$ denote a divergent series of positive numbers, such that the limit of l_n, as n is indefinitely increased, is zero. Let a be a positive number < 1, and let $n_1, n_2, \dots n_\iota, \dots$ be a sequence of increasing positive integers. It is possible to choose the divergent series so that each of the ratios $l_{n_2}/l_{n_1}, l_{n_3}/l_{n_2}, \dots$ is $< a$; if this be done, the series $\sum_{\iota=1}^{\infty} l_{n_\iota}$ is convergent, its sum being $< \frac{l_{n_1}}{1-a}$. Each of the series obtained from $\sum_{\iota=1}^{\infty} l_{n_\iota}$, by leaving out a finite number of terms, is also convergent. The convergent series, so obtained, form an unenumerable set: for they are obtained by multiplying the terms of the series $\sum_{\iota=1}^{\infty} l_{n_\iota}$, each either by 0, or by 1; and thus there is a series corresponding to each fractional number expressed in the dyad scale. Corresponding to each convergent series there is a divergent series which consists of $l_0 + l_1 + \dots + l_n + \dots$, with the convergent series removed from it. We obtain in this manner an unenumerable set of divergent series. The convergent and divergent series, each of which consists of terms of $l_0 + l_1 + \dots + l_n + \dots$, may now be correlated with the numbers of the continuum $(0, 1)$. Let these numbers be expressed in the dyad scale, in the form $\cdot a_1 a_2 a_3 \dots$, where every a is 0, or 1, and the case in which every figure is zero, after some place, is excluded. To one of the series $l_p + l_q + l_r + \dots$, we may take that number to correspond in which $a_p, a_q, a_r, \dots$ are all 1, and the remaining digits 0. The points of $(0, 1)$ are thus divided into two classes, one of these consisting of all the numbers which correspond to convergent series, and the other of those corresponding to divergent series.

ORDINARY INNER LIMITING SETS

97. The inner limiting set† of a sequence of open sets, each one of which contains the next, is said to be *an ordinary inner limiting set*.

The outer limiting set of a sequence of closed sets, each of which is contained in the next, is said to be *an ordinary outer limiting set*.

---

* *Crelle's Journal*, vol. CXVIII (1897), p. 29.

† The term inner limiting set is due to W. H. Young, who has investigated the properties of such sets; see *Leipz. Ber.* 1903, p. 287; for further properties see also his papers in the *Proc. Lond. Math. Soc.* (2), vol. I (1903), p. 212, and (2), vol. III (1905), p. 372, where the theory is extended to p-dimensional sets.

Thus an ordinary inner limiting set is the complement of an ordinary outer limiting set.

In the case of a linear set of points contained in a given interval (a, b), to which the case of a linear set in an unbounded interval may be reduced by the method of correspondence, an ordinary inner limiting set is the set of points common to a sequence $\{\Delta_n\}$, where Δ_n is, for each value of n, a set of open intervals contained in (a, b).

In the particular case in which all the sets Δ_n are everywhere dense in (a, b), the complementary closed sets $C(\Delta_n)$ are all non-dense. In this case the ordinary inner limiting set is of the second category, and is of that species which is termed (§ 95) a residual set.

The case of linear sets will here be considered in detail.

The following theorem will be established:

Every ordinary (linear) limiting set is either enumerable, or it has the power of the continuum.

If the sets Δ_n are all everywhere dense in (a, b), the inner limiting set, being a residual set (see § 100), has the cardinal number of the continuum, in accordance with the theorem of § 95.

Next, let us suppose that the sets $\{\Delta_n\}$ are not all of them everywhere dense in (a, b); and let us suppose further that the inner limiting set contains a part H which is dense in itself, so that H' is perfect. The set H' may be placed in correspondence with points of the interval $(0, 1)$, so that the order of corresponding points in H' and in the interval $(0, 1)$ is always the same. To each point in $(0, 1)$ there corresponds a single point of H', except that the end-points of an interval contiguous to H' correspond to a single point in $(0, 1)$. The points of H correspond to the points of a set H_1 everywhere dense in $(0, 1)$; and those intervals of Δ_n that contain points of H correspond to a set of intervals Δ_n' everywhere dense in the interval $(0, 1)$. The set H therefore corresponds to a residual set in $(0, 1)$; and therefore H has the cardinal number of the continuum.

The only sets of points which contain no components that are dense in themselves are enumerable sets, and therefore the cardinal number of an ordinary inner limiting set is c or a.

It has thus been established that:

An ordinary inner limiting set has the power of the continuum if it contains a set that is dense in itself; and if it contains no such component it is enumerable.

An enumerable set which contains a component that is dense in itself cannot be an ordinary inner limiting set.

98. If any linear set G is given, we may suppose a point x. of G, to be in the interior of each interval of a sequence $\delta_1(x), \delta_2(x), \dots \delta_n(x), \dots$ of intervals, each one of which contains the next; and this will be the case for each point x in G. Let d_n be the upper boundary of the lengths of the intervals $\delta_n(x)$, taken for every point x in G; and suppose the intervals so determined that the numbers $d_1, d_2, \dots d_n, \dots$ form a sequence which converges to zero. Let Δ_n be the set of open non-overlapping intervals which have the same interior points as the set of intervals $\{\delta_n(x)\}$. We have then a sequence of sets $\Delta_1, \Delta_2, \dots \Delta_n, \dots$ of open intervals, which will define an ordinary inner limiting set, to which all the points of G must belong; but it may also include limiting points of G that do not belong to G. For it is clear that any point of the inner limiting set is at a distance from some point of G_n that does not exceed d_n; hence any point of the inner limiting set which is not a point of G_n is the limiting point of a sequence of points of G_n, since $d_n \sim 0$, as $n \sim \infty$.

The set of points of G' which do not belong to G, but which belong to the inner limiting set defined by the sequence $\{\Delta_n\}$, will depend in general on the mode in which the intervals of the sets Δ_n are defined. In order that the given set G may be an ordinary inner limiting set it must be possible so to define the sets Δ_n that no point G' that does not belong to G is in Δ_n for every value of n.

Some criteria will be here given for deciding in respect of a given set whether this is possible or not, that is, whether the set is an ordinary inner limiting set or not.

A point which is in Δ_m, but not in the sets $\Delta_{m+1}, \Delta_{m+2}, \dots$, is said to be *shed* from the sequence $\{\Delta_n\}$ at the index m.

If all the intervals $\{\delta_n(x)\}$ be taken to be of equal length $2c_n$, with the point x in the centre of $\delta_n(x)$, where $c_n \sim 0$, as $n \sim \infty$, then every limiting point of G belongs to the ordinary inner limiting set defined by the sequence of intervals.

For, however small c_n may be, there are points of G whose distance from a limiting point p is less than c_n.

EXAMPLE

The following example, given by Borel*, drew attention to the fact that a sequence of sets of intervals constructed as above may define an inner limiting set that contains points other than those of the given set.

Let us suppose that each rational point $\frac{p}{q}$ in the interval $(0, 1)$ is enclosed in the interval $\left(\frac{p}{q} - \frac{\lambda}{q^3}, \frac{p}{q} + \frac{\lambda}{q^3}\right)$, where λ has the same value for all the points. In this manner the rational points are enclosed in a set of overlapping intervals, whose sum is less than $\lambda\Sigma(q-1)\frac{2}{q^3}$, or than $2\lambda\Sigma\frac{1}{q^2}$, which can be made as small as we please by choosing λ small enough. The equivalent set of non-overlapping intervals defines, by means of the end-points and their limits, a closed set $\{q_1\}$, such that for any point of the set $\left|\frac{p}{q}-q_1\right| \geq \frac{\lambda}{q^3}$, for all points $\frac{p}{q}$.

---

* *Leçons sur la théorie des fonctions*, p. 44.

Now consider the set of points defined by

$$x=\frac{a_1}{10}+\frac{a_2}{10^{2!}}+\frac{a_3}{10^{3!}}+\dots+\frac{a_n}{10^{n!}}+\dots,$$

where each a is $\leqq 9$, and the a's are such that an infinite number of them are different from zero. It has been shewn by Liouville that these numbers x are transcendental. Let

$$\frac{p}{q}=\frac{a_1}{10}+\frac{a_2}{10^{2!}}+\dots+\frac{a_n}{10^{n!}},\ \text{thus } q=10^{n!},$$

then
$$x-\frac{p}{q}=\frac{a_{n+1}}{10^{(n+1)!}}+\dots<\frac{1}{q^n}\left(\frac{a_{n+1}}{q}+\dots\right)<\frac{1}{q^n}.$$

It follows that, if x is one of the above transcendental numbers, whatever value λ may have, it is interior to an interval $\left(\frac{p}{q}-\frac{\lambda}{q^3}, \frac{p}{q}+\frac{\lambda}{q^3}\right)$. For suppose $q=10^{n!}$; then $\left|\frac{p}{q}-x\right|<\frac{1}{q^n}<\frac{\lambda}{q^3}$, provided $\lambda \geqq \frac{1}{10^{(n-3)n!}}$; and, however small λ may be, values of n can be found for which this inequality is satisfied. Therefore rational points $\frac{p}{q}$ can be found, however small λ may be, such that x lies within the intervals $\left(\frac{p}{q}-\frac{\lambda}{q^3}, \frac{p}{q}+\frac{\lambda}{q^3}\right)$. It thus appears that, besides the original points $\frac{p}{q}$ which the intervals are drawn to enclose, there are other points which lie inside the intervals, for all values of λ, when λ is diminished indefinitely

99. *Those limiting points of an enumerable set of points P that do not belong to P form an ordinary inner limiting set.*

Let $p_1, p_2, \dots p_n, \dots$ denote the points of P, and let H denote the set of those limiting points of P which do not belong to P.

The points of H can be enclosed in a set Δ_1 of open intervals that do not include the point p_1. Interior to Δ_1, a set of open intervals Δ_2 can be defined which include all the points of H, but not the point p_2; and so on. The sequence $\Delta_1, \Delta_2, \dots$ has for its inner limiting set all the points of H, but none of the points $p_1, p_2, \dots$; and H can have no limiting points that do not belong either to itself or to P. Hence H is an ordinary inner limiting set.

Every isolated set is an ordinary inner limiting set.

For each point x, of an isolated set, may be enclosed in an interval of a sequence $\delta_1(x), \delta_2(x), \dots \delta_n(x), \dots$, each interval containing the next, and so that $\delta_n(x)$ converges to zero; and such that $\delta_1(x)$ contains no point of the given set other than x. The set Δ_n may then consist of the intervals $\delta_n(x)$ for every point x of the set. It is then clear that the only points interior to Δ_n, for every value of n, are the points of the given isolated set, which is therefore an ordinary inner limiting set.

Every closed set is an ordinary inner limiting set.

Enclose each point P, of a closed set G, in an interval of length 2ρ, with P as its middle point; any part of such an interval that is not in (a, b) may be disregarded. We have thus a set of overlapping intervals which contain all the points of G within them. Consider the equivalent set of non-overlapping intervals (§ 71). We have thus a set Δ, of non-overlapping intervals, each of which is of length $\geq \rho$; and therefore Δ must be a finite set, which contains all the points of G. Let ρ have all the values in a sequence of diminishing numbers that converges to zero; the corresponding sequence of finite sets of intervals $\Delta_1, \Delta_2, \ldots \Delta_n, \ldots$ is such that each contains the next, and every interval includes points of G. The inner limiting set which they define must be the set G, since G contains all its limiting points.

100. *The necessary and sufficient condition that an enumerable set P should be an ordinary inner limiting set is that P contains no component that is dense in itself.*

That the condition is necessary has been shewn above, since every inner limiting set that has a component that is dense in itself has the power of the continuum.

To prove the sufficiency of the theorem, we employ the mode of analysis of a set, given in § 91, into adherences and coherences.

Any enumerable set P can be resolved into a sum of sets

$$P_1 + P_2 + \ldots + P_\beta + Q_\beta,$$

where $P_1, P_2, \ldots P_\beta$ are all isolated sets, and Q_β is a component of a perfect set G_β: β denotes a number of the first, or of the second, class. When P contains no component that is dense in itself, the decomposition of P ceases at, or before, some definite number α of the first, or of the second, class, when there is no set G_α.

It has thus been shewn that, when P contains no component that is dense in itself, it can be resolved into a finite, or enumerably infinite, set of inner limiting sets, of which there may, or may not, be a last set. Let P_γ be one of the components into which P has been resolved, γ denoting a number of the first or second class. We now fix on a sequence of sets of intervals enclosing the points of P_γ, such that all the intervals are interior to the intervals complementary to G_γ; then the set

$$P_{\gamma+1} + P_{\gamma+2} + \ldots,$$

which is contained in G_γ, has no limiting points in any of the intervals which enclose the points of P_γ, for all its limiting points must be in G_γ. The sequence of sets of intervals having thus been fixed for every P_γ, we can now shew that each limiting point p, of P, which does not belong to P, is shed from the whole sequence of sets of intervals, at a definite index.

The point p is either a limiting point of P_1, belonging to L_1, the set of the isolated points of P', or is contained in G_1. In the former case it is shed from the intervals enclosing P_1 at a definite index; and, not being a limiting point of $P_2 + P_3 + \dots$, it is shed from the intervals enclosing the points of that set, at a definite index; consequently it is shed from the intervals enclosing P, at a definite index, the greater of the two former ones. In the latter case, unless p is in G_2 or in P_2', it is not a limiting point of $P_2 + P_3 + \dots$, and does not come into any of the intervals enclosing the points of P_1; it is therefore shed at a definite index. If p belongs to $G_1, G_2, \dots$ and to every G before G_α, but is not in G_α, it may be a point of P_α'. In that case it is not a limiting point of the set

$$P_{\alpha+1} + P_{\alpha+2} + \dots,$$

and does not come into the interior of any of the intervals which enclose the points of $P_1, P_2, \dots$, or any P with index less than α. It is therefore shed, at a definite index, from the sequence of sets of intervals enclosing the points of P. The theorem has thus been established.

A corollary to the above proof is that every enumerable set is the sum of an ordinary inner limiting set, and of a set which is dense in itself.

In particular* it follows that.

Every reducible set is an ordinary inner limiting set.

For the derivative of a reducible set being enumerable, the given set can contain no component that is dense in itself, since the derivative of such component would be perfect, which is impossible when the derivative is enumerable.

Since an ordinary inner limiting set that is everywhere dense consists of the points that are common to a sequence $\{\Delta_n\}$ of sets Δ_n, of non-overlapping open intervals which are everywhere dense, and since the points that do not belong to Δ_n form a non-dense closed set, it follows that *an everywhere dense ordinary inner limiting set is a residual set* (§ 95): the complementary set, of the first category, being the outer limiting set of a sequence of non-dense closed sets.

101. Any unenumerable set can, in accordance with the result of § 91, be expressed in the form $P = U + V_a + \Sigma V_x + V_c$; and we observe that, if V_c is absent, the necessary and sufficient conditions that P may be an inner limiting set are that V_a and ΣV_x should both be absent; this follows from the preceding results.

If V_c exists, we observe that no point of $U + V_a + \Sigma V_x$ can be a limiting point of V_c; for any limiting point of V_c must be a point of degree c in the set P. If V_c is everywhere dense in (a, b), it follows that $U + V_a + \Sigma V_x$ is absent. The set V_c may be non-dense in (a, b), or it may be dense in some parts of (a, b), and non-dense in other parts.

---

* See Hobson, *Proc. Lond. Math Soc.* (2), vol. II (1904), p. 316 Another proof of the general theorem, applicable to any number of dimensions, has been given by Brouwer, *Proc. Roy. Soc. Amsterdam*, vol. XVIII (1) (1915), p. 48. See also vol. XX (1917) for further information on the subject.

It will be shewn that V_c is in general made up of a part which is non-dense in (a, b), and of a finite, or indefinitely great, number of parts each of which is everywhere dense in a particular interval in which it lies. Suppose that an interval (α, β) can be found in which V_c is everywhere dense; and let x be a point in (a, b) such that $x \geq \beta$. Then those values of x for which V_c is everywhere dense in (α, x), together with those values for which this is not the case, define a section of all the numbers of the continuum (β, b); and this section defines a number $\beta_1 \geq \beta$. Similarly, we may assign a number $\alpha_1 \leq \alpha$, so that (α_1, β_1) is the greatest interval containing (α, β) which is such that V_c is everywhere dense in it. If, in the parts of (a, b) external to (α_1, β_1), the set V_c is dense in any interval, then we proceed to fix the greatest interval for which it is everywhere dense. In this manner we obtain a finite, or enumerably infinite, set of detached intervals contained in (a, b), in each of which V_c is everywhere dense, and the remainder of (a, b) may consist of a set of detached intervals and a set of points. In this remainder the points of V_c form a non-dense set.

No point of $U + V_a + \Sigma V_x$ can be in an interval (α_1, β_1) in which V_c is everywhere dense. If $\overline{V}_c$ is the part of V_c which is non-dense in (a, b), every point of $U + V_a + \Sigma V_x$ must lie in one of the intervals complementary to the perfect set $\overline{V}_c'$. It is to be observed that in $\overline{V}_c$ are included the end-points of the intervals (α_1, β_1), in case those end-points belong to V_c.

In order that P may be an inner limiting set, it is necessary that the part of $U + V_a + \Sigma V_x$ which is in each interval complementary to $\overline{V}_c'$ should be an inner limiting set; and this cannot be the case unless V_a and ΣV_x are absent.

It has thus been shewn that:

In order that an unenumerable set of points may be an inner limiting set, it is necessary that the set should contain no points whose degrees in the set are other than 0, *a, or c, and that it should contain no component which is dense in itself, and whose points are of degree a in the set.*

The determination of the necessary and sufficient conditions that any given unenumerable set of points, however defined, may be an ordinary inner limiting set, has now been reduced to the problem of determining the criteria for the case of a set which is dense in itself, and all the points of which are of degree c in the set. The case in which the latter set is non-dense in its domain may be reduced, by the method of correspondence, to that in which it is everywhere dense; and the problem is therefore reducible to that of determining the conditions under which a given everywhere dense set of points, all of degree c in the set, is a residual set, of the kind described in § 100. No investigation of all the possible types of such sets has yet been carried out, and therefore the problem remains as yet unsolved.

PLANE SETS OF POINTS

102. It has already been shewn in § 85, in the case of closed sets, that the descriptive properties of sets of points in two or more dimensions are in some respects less simple than those of linear sets. The properties of plane sets of points are of special importance in the theory of functions of a complex variable, but they do not differ in any essential respect from the properties of sets in three or more dimensions. The definitions of closed sets, open sets, frontiers, etc., given in § 55, for such sets, are identical with those for linear sets, but an account will here be given of the most important descriptive properties of sets in more dimensions than one. Although the proofs of the theorems are, for simplicity of language, given for plane sets only, they may be extended to the case of sets in more dimensions than two, without material alteration

Each point (x, y), of a plane set, is defined by the two real numbers which are the rectangular Cartesian coordinates of the point. As has already been pointed out in § 53, a correspondence may be established between the points in a finite cell, or rectangle, and the points in unbounded space, of such a character that the descriptive properties of sets of points are unaltered by the transformation; a special convention being introduced concerning the points on the boundary of the finite rectangle, or cell. It makes no essential difference, in the properties which will here be considered, whether the closed or open sets are considered as bounded, that is as contained in a finite open, or closed, rectangle, or whether they are in the unbounded plane space, provided that, when necessary, an adjunct boundary at infinity is postulated which corresponds to the boundary of a finite rectangle.

103. The frontier of a set of points G has been defined, in § 55, as the set of points each of which belongs to one of the sets G, $C(G)$, and is a limiting point of the other set. It will be shewn* that:

If the complementary set of G exists, then the frontier of G and $C(G)$ exists, and is a closed set.

Let P be any point of G, and P' a point of $C(G)$, and consider those points of G that are on the linear segment PP', *i.e.* those points whose coordinates are $\frac{x+kx'}{1+k}$, $\frac{y+ky'}{1+k}$, where P is (x, y), and P' is (x', y'), and k is a positive number, or zero. The linear set of points of G, on

---

* Jordan, *Cours d'Analyse*, vol. I, p. 20.

PP', has, in accordance with the theorem of § 47, an upper boundary Q. This point Q, which may coincide with P, is a point of the frontier of G and $C(G)$; for if Q is a point of G, it is also a limiting point of $C(G)$, and if it is a point of $C(G)$, it is a limiting point of G. Therefore, if $C(G)$ exists, there is always a frontier of G and $C(G)$. Moreover, let

$$Q_1, Q_2, \ldots Q_n, \ldots$$

be a convergent sequence of points of this frontier. Denoting by $\bar{Q}$ the limiting point of the sequence, $\bar{Q}$ is itself a point of the frontier; for in the set $\{Q_n\}$ there is contained a convergent sequence of points all of which belong to G, or else such a sequence of which all points belong to $C(G)$. If these points all belong to $C(G)$, and consequently to G', then $\bar{Q}$ belongs to the closed sets G' and $\{C(G)\}'$; if they all belong to G, and consequently to $\{C(G)\}'$, then $\bar{Q}$ belongs to G' and to $\{C(G)\}'$. In either case $\bar{Q}$ is a point of the frontier; and thus, since every limiting point of the frontier belongs to it, the frontier is a closed set.

If all points of G belong to its frontier, G has no interior points.

104. The distance* of two points P, or (x, y), and Q, or (x', y'), has been defined in § 50 to be the number $\{(x - x')^2 + (y - y')^2\}^{\frac{1}{2}}$; and this may be denoted by PP'.

If P is a point of a set G_1, and P' a point of another set G_2, the distances PP', when every such pair of points is contemplated, form an aggregate of numbers which have a lower boundary. In case this lower boundary is a positive number $d(G_1, G_2)$ (> 0), the sets G_1 and G_2 are said to be *detached* from one another. The number $d(G_1, G_2)$ is said to be the *distance* between the two sets G_1, G_2.

If one set G_2 consists of a single point p, $d(p, G_1)$, or simply d, is said to be the distance of the point p from the set G_1.

If two closed sets, G_1, G_2 are detached from one another, there exists at least one pair of points P, P', belonging respectively to the two sets, such that the distance PP' is equal to the distance of the sets from one another.

Let d denote the distance $d(G_1, G_2)$, of the closed sets from one another; and let $\epsilon_1, \epsilon_2, \ldots \epsilon_n, \ldots$ be a sequence of decreasing positive numbers that converges to zero. For each value of n, a pair of points P_n, or (x_n, y_n), and P_n', or (x_n', y_n'), belonging to G_1, G_2 respectively, can be determined such that $P_nP_n'^2 < d^2 + \epsilon_n$. A unique point p_n, or

$$(x_n, y_n, x_n', y_n'),$$

exists in the four-dimensional continuum, corresponding to each pair of points P_n, P_n'. The set of points $p_1, p_2, \ldots p_n, \ldots$ has at least one limiting point (x, y, x', y'); let P, P' denote the two points (x, y), (x', y') in the plane space. It will be shewn that P and P' belong to G_1, G_2 respectively, and that PP' is equal to d. An integer m can be so determined that $x - x_n$, $y - y_n$, $x' - x_n'$, $y' - y_n'$ are all numerically less than an arbitrarily chosen positive number η, provided $n \geqq m$. It follows that P is a limiting point of the sequence $\{P_n\}$, and that P' is a limiting point of the sequence $\{P_n'\}$; and thus that P and P' belong to G_1, G_2 respectively, since these sets are closed. We have, further,

---

* Instead of the distance so defined, Jordan employs in this connection the "écart," defined by $|x - x'| + |y - y'|$. This makes no difference in the developments.

$$| x - x' | \leqq | x - x_n | + | x_n - x_n' | + | x_n' - x' | \leqq 2\eta + | x_n - x_n' |,$$

and similarly $| y - y' | \leqq 2\eta + | y_n - y_n' |$, for $n \geqq m$. From these inequalities we see that $(x - x')^2 + (y - y')^2 < 8\eta^2 + A\eta + P_nP_n'^2$, where A is some fixed number; hence $PP'^2 < 8\eta^2 + A\eta + d^2 + \epsilon_n$. Since η and ϵ_n are both arbitrarily small, it follows that $PP'^2 \leqq d^2$, and thus that PP'^2, which is certainly not less than d^2, must be equal to d^2. The theorem has now been established. For the choice* of the infinite set of pairs of points P_n, P_n', the multiplicative axiom is required.

105. A closed set of points is said to be *connex*, or single sheeted, when it is not the sum of two or more closed sets.

It should be observed that the expression "connex" has been employed in § 41 in a somewhat different sense.

A connex closed set which does not consist of a single point is a perfect set.

For an isolated point of a closed set G can be considered as a set detached from the closed set consisting of all the remaining points of G; and hence, if an isolated point existed, G could not be connex.

If P, P' are any two points of a connex perfect set G, and ϵ be any assigned positive number, a finite number of points $p_1, p_2, \dots p_n$, of the set, can be so determined that $Pp_1, p_1p_2, p_2p_3, \dots p_{n-1}p_n, p_nP'$ are all $\leqq \epsilon$. Conversely, if this condition holds for each pair of points P, P' of G, and for every positive number ϵ (n depending on ϵ), then G is connex.

The condition stated in the theorem is sufficient to ensure the connexity of G. For if G be the sum of two closed sets G_1, G_2, the distance between which is d (> 0), we may choose ϵ to be $< d$. If P is a point of G_1 and P' a point of G_2, and p_1 is a point of G such that $Pp_1 \leqq \epsilon$, the point p_1 belongs to G_1; again if p_2 be a point of G such that $p_1p_2 \leqq \epsilon$, the point p_2 belongs to G_1; and so on. Since p_n belongs to G_1, whatever finite value n may have, it is impossible that $p_nP' \leqq \epsilon$, because $p_nP' \geqq d$. Again, the condition is a necessary one. For, let us suppose that, for some value of ϵ, the condition is not satisfied for every pair of points. If P belong to such a pair, the set G may be divided into two parts G_1 and G_2; where G_1 is such that, for each point P' belonging to it, a definite set of points of G_1, viz. $p_1, p_2, \dots p_n$, exists, such that $Pp_1, p_1p_2, \dots p_nP'$ are all $\leqq \epsilon$;

---

* See Chap. IV, § 197 *et seq.*

and G_2 is such that, for each point of it, this condition is not satisfied. It can be shewn that G_1, G_2 are both closed sets, and that the distance between them is $> \epsilon$. For, if p is a limiting point of G_1, it belongs either to G_1 or to G_2, and since there are points p_n, of G_1, such that $pp_n < \epsilon$, the point p clearly belongs to G_1: therefore G_1 is a closed set. Again, if q is a limiting point of G_2, it cannot belong to G_1; for a point P', of G_2, can be found such that $qP' < \epsilon$; hence, if q belonged to G_1 so also would P'; thus G_2 is closed. It is clear that no pair of points of G_1, G_2 can exist, of which the distance is $\leq \epsilon$; hence, for these sets, $d > \epsilon$. It has thus been shewn that if, for any value of ϵ, the condition in the theorem is not satisfied, G can be divided into two detached closed sets, and it is therefore not connex.

106. An open set, defined in § 55 as one such that every point is an interior point, is also frequently called an open domain.

When the open domain is such that any two points P, P' of it may be joined by means of a finite set of linear intervals Pp_1, p_1p_2, ... p_nP', such that all the points of all these closed intervals belong to the domain, it is said to be a *connex open domain.*

A connex open domain is also called a *Weierstrassian domain*, or a *continuum*. It may be either bounded or unbounded. The terms domain and region are also used to denote a connex open domain together with some, or all, of its frontier points.

If, to a connex open domain, all its frontier points be added, it becomes a closed set, connex in accordance with the definition of § 105 But a closed connex set does not necessarily become a single connex open domain when all its frontier points are removed from it. Thus, for example, the set of all the points interior to, or on the circumferences of, two circles that touch each other externally, is a closed set, but the interior points of the circles do not form a single connex open set, but two such sets.

Each point of an open domain is such that a circle exists with the point as centre, and with all its interior points in the domain, but with one or more points of the frontier on its circumference.

For the point P has a definite distance d_P from the closed frontier of the open domain, and in virtue of the theorem of § 104 there is one point at least of the frontier on the circumference of the circle with centre P and radius d_P.

Every open set of points is the sum of a finite, or an enumerably infinite, set of connex open domains.

If P be any point of the given open set O, let the component O_1, of O, be such that every point Q, of O_1, satisfies the condition that P and Q can be joined by a set Pp_1, p_1p_2, ..., p_nQ of straight lines of which all the points are points of O. The set O_1 is clearly an open set, for a point Q of it has a neighbourhood, of which every point q can be joined in the above manner to P; and thus Q is an interior point of O_1. Therefore any point P, of O, is a point of a connex open domain which is a component of O. It will be shewn that the number of such connex domains contained in O is enumerably infinite, or finite. Apply to the finite, or infinite, rectangle, in which O is contained, a system of closed nets $\{D_n\}$. There is a smallest integer n_1, such that D_{n_1} has one or more meshes interior to O; there is a smallest integer n_2 $(> n_1)$, such that those meshes of D_{n_1} which are not interior to O contain one or more meshes of D_{n_2} that are interior to O; and so on. We obtain in this manner a sequence of sets of meshes $d_{n_1}, d_{n_2}, \ldots$ such that each mesh of the sequence is interior to O, and such that no mesh of the sequence contains any other such mesh. Every such component of O, as O_1, contains meshes of the sequence; and since the set of meshes in the sequence is enumerable, it follows that the aggregate of all components of O that consist of connex open sets is enumerable, whether finite or not. It follows from this theorem that *every closed set of points has for its complement, relatively to a finite, or infinite, cell, a finite, or an enumerably infinite, set of connex open domains. If the set of connex open domains is everywhere dense in a closed domain, the closed set is non-dense in that domain.*

These connex open domains are the true analogues, for plane closed sets, of the open intervals contiguous to a linear closed set of points, and they may therefore be said to be the open connex domains *contiguous* to the closed set. In § 86, attention has been confined to domains of a particular type, viz. cells, and it has there been shewn that there exists, in general, no unique set of cells which have properties in relation to a given closed set fully analogous to the properties of the open intervals contiguous to a linear closed set. If, to a non-dense closed set, there be added all the interior points of some of the contiguous domains, we obtain a closed set of the most general type. A closed set will be perfect, if no two of the contiguous domains abut on one another, or also if they abut on one another in such a way that the points common to the two boundaries form a perfect set.

107. *A bounded connex domain together with its frontier points forms a connex closed set.*

Since the frontier is closed, it is easily seen that the set formed by adding all the points on the frontier of the bounded connex domain is closed.

If P, Q are both interior points of the closed set, they are both points of the bounded connex domain, and thus satisfy the requisite conditions as to the mode in which they may be joined. If P, Q are points on the frontier, interior points P', Q' may be so determined that PP', QQ' are both less than ϵ; and then P', Q' may be joined by a series of straight lines $P'p_1, p_1p_2, \ldots, p_nQ'$ each of length less than ϵ, and such that every point of all of them is an interior point of the connex domain. It thus follows that P, Q may be joined so that the requisite condition is satisfied. The theorem will also hold good for an unbounded domain, provided those points at infinity which are boundary points are adjoined.

The boundary of a connex domain, although closed, is not necessarily itself connex. For example, if the connex domain consists of the points between two concentric circles, the boundary consists of the two circumferences, and is not connex

The Heine-Borel theorem in its generalized form, proved in § 74, has already been emancipated, in § 75, from the condition that the set to which the points of the closed set are interior is necessarily a set of cells. It has in fact been shewn in § 75 that a set of open domains may be employed instead of a set of cells. The theorem may then be stated in the following form:

If every point of a (bounded) closed set G belongs to one or more of the connex open domains of a set D (not necessarily enumerable) of such domains, there exists a finite set of connex open domains belonging to D such that every point of G belongs to one at least of the domains of the finite set.

In a similar manner it is shewn that, in the last theorem of § 75, applicable to any set of points G, open connex domains may be employed instead of cells. We thus have the following theorem:

If each point of a given set of points belongs to a definite connex open domain corresponding to that point, an enumerable set of these connex open domains can be so determined that all the points of the set belong to one or more of the domains of that enumerable set.

If the enumerable set of connex open domains be $\{D_n\}$, it may be shewn that, if H be any closed set contained in G, all the points of H belong to one or more of the domains of a finite set which is a component of $\{D_n\}$.

For, since each point of the closed set is interior to one or more of the domains $\{D_n\}$, the result follows from the generalized Heine-Borel theorem.

108. The case when the set is a bounded open set O, in which, to each point of O, there corresponds a domain which consists of the interior of a circle with the point as centre, is of special importance in connection with the theory of a complex variable.

Let d_P be the distance of the point P, of a bounded open set O, from the frontier of the set. To the point P, the interior of the circle of radius d_P, and centre P, may be taken to correspond. In accordance with the theorem of § 107, every point of the open set is interior to one at least of an enumerable set $\{C_n\}$ of these circles. Each of the circles has on its circumference one or more points of the frontier of the open set. If the radius of the circle with centre P be θd_P, instead of d_P, where θ is some fixed number (< 1), an enumerable set of these circles $\{\bar{C}_n\}$ exists, such that each point of the open set is interior to one or more of the circles $\{\bar{C}_n\}$, and all the points on the circumferences of these circles are points of the open set O.

The results may thus be stated as follows:

If O be any bounded open set, there exists (1), *an enumerable set of circles $\{C_n\}$ such that each point of O is an interior point of one at least of the circles, and such that every interior point of the circles is a point of O, and the circumference of each of the circles contains one or more points on the frontier of O; and* (2), *there exists an enumerable set of circles $\{\bar{C}_n\}$ such that each point of O is interior to one or more of the circles, and every point on the circumference, or interior to, any of the circles, is a point of O.*

If O consist of a single connex open domain, any two points P, Q, of O, can be joined by means of a finite number of the circles of the set $\{C_n\}$, or also by a finite number of the circles of $\{\bar{C}_n\}$; so that P, Q are interior to circles of the finite set, and each circle overlaps the next.

For P and Q may be joined by a finite set of segments of straight lines all of which segments are points of O. This broken line joining P and Q constitutes a closed set H, which, in accordance with the last remark in § 107, is such that each point of H is interior to a finite set of circles belonging either to $\{C_n\}$, or to $\{\bar{C}_n\}$. Whichever of these sets of circles be employed, we have a chain of circles such that P and Q are each interior to one of them, and such that each circle of the chain overlaps the next one.

109. The following theorem is the analogue, for a plane set of points, of the theorem proved in § 71 for an open linear set.

Having given a set of overlapping open domains, there exists a set of non-overlapping open domains such that an interior point of a domain of either set is an interior point of a domain of the other set.

The proof of the theorem is the same as that in § 71; it being shewn that every point interior to one or more of the given domains is an interior point of the set of all such points. The set of all points interior to one or more of the given domains is open, and consequently consists of a finite, or an enumerable set of connex open domains.

Every open domain O is the sum of an enumerable set of closed cells which do not overlap, though any two of them may have a portion of their boundaries in common.

We may prove the theorem for the case in which the open domain is bounded, as it can then be extended to the case of an unbounded open domain, by the method of correspondence.

Apply a system of nets, with closed meshes, to a fundamental cell which contains the open domain. There is a smallest value n_1, of n, such that D_{n_1} has one or more meshes which, being completely closed, are interior to O. Let these meshes, arranged in order of rank, be denoted by $\bar{D}_{n_1}$. There is a smallest value n_2, of n, $(> n_1)$ so that D_{n_2} contains meshes not contained in $\bar{D}_{n_1}$, each of which is interior to O. Proceeding in this manner, we have a sequence of finite sets of meshes $\bar{D}_{n_1}, \bar{D}_{n_2}, \ldots$, arranged in order, such that each mesh is interior to O, and no two of the meshes overlap one another.

Any point of O is contained in each of a unique sequence of meshes $d_1, d_2, d_3, \ldots$ belonging to $D_1, D_2, \ldots$ respectively. These are the meshes in which the point would have been contained if the nets had been half closed instead of completely closed. For all sufficiently large values of n, d_n is interior to O. The smallest such value of n must be one of the numbers $n_1, n_2, \ldots$; say n_{p_1}. Thus the point P, of O, is in a mesh $d_{n_{p_1}}$ which belongs to $\bar{D}_{n_{p_1}}$. It is thus proved that all the points of O are points of the sequence $\bar{D}_{n_1}, \bar{D}_{n_2}, \ldots$.

An open set is the outer limiting set of a sequence of closed sets

First let the open set O be bounded, and let H be its boundary. Let d be a fixed number, and consider the set of points G_d which consists of all points of O at a distance from H that is $\geq d$. This set G_d, which certainly exists provided d be sufficiently small, is a closed set. For let $\{P_n\}$ be a convergent sequence of sets of points of G_d that has P for its limiting point.

The distance of P from H cannot be less than d, for if it were, for a sufficiently great value of n, P_n would be at a distance from H less than d, which is not the case. If d have the values in a sequence $d_1, d_2, \ldots d_n, \ldots$ which steadily converges to zero, the closed sets $G_{d_1}, G_{d_2}, \ldots G_{d_n}, \ldots$ are such that each is contained in the next, and any point of O belongs to all of the closed sets, from and after some value of n. Thus O is the outer limiting set of the sequence $\{G_{d_n}\}$ of closed sets.

In case the open set is unbounded, the theorem may be deduced from the case in which it is bounded. In this case the closed sets may have parts of their frontiers at infinity.

The theorem may also be proved by considering the set of circles $\{\bar{C}_n\}$ such that every point of O is an interior point of one or more of the circles, and that every point in, or on the circumference of, each circle belongs to O. For we may consider the closed set $M(\bar{C}_1, \bar{C}_2, \ldots \bar{C}_n) \equiv G_n$; then O is the outer limiting set of the sequence $\{G_n\}$ of closed sets.

A closed set is the inner limiting set of a sequence of open sets.

This is at once deduced from the last theorem by considering the complementary sets.

110. *Having given a non-finite family of sets of points, there exists at least one point such that, in its arbitrarily small neighbourhood, there are points belonging to an infinite number of sets of the given family. Such a point may be an adjoined point at infinity.*

In the case in which each of the sets of points consists of a single point this theorem reduces to that of the existence of the limiting point of a set of points; and thus the theorem may be considered to be a generalization of the fundamental theorem of § 52.

In case all the sets of the family are contained in a finite fundamental cell, the theorem is proved exactly as that in § 52. A system of nets being applied to the fundamental cell, a sequence of meshes $d_1, d_2, \ldots$ each containing the next, and belonging to the successive nets, is obtained, and each of which contains points belonging to an infinite number of the sets of the given family. The point defined by the sequence $\{d_n\}$ is a point which satisfies the condition of the theorem. In case there is no finite fundamental cell, the theorem is proved by the method of correspondence with a family of sets in a finite cell. In this case the point determined may be an adjoined point at infinity.

It is easily seen that all the points that satisfy the condition of the theorem form a closed set; the points at infinity, if any, being included in the set.

In case all the sets are open domains, we have the following theorem:

Having given a non-finite (enumerable or unenumerable) family of open domains, there exists at least one point such that, in its arbitrarily small neighbourhood, there are contained domains which are portions of an infinite number of the domains of the given family. Such a point may be an adjoined point at infinity. The set of all such points is a closed set.

111. The theorems in § 82 can be extended to the case of a set of two (or more) dimensions.

Instead of intervals $\delta_1, \delta_2, \ldots \delta_n, \ldots$, each containing the next, and converging to the point p, we take a convergent set of cells. Instead of intervals contiguous to the perfect, or the closed, set P_Ω, we take the set of contiguous domains. The procedure in the proofs is then otherwise unaltered.

In case the sets $P_2, P_3, \ldots P_\alpha, \ldots$ consist of the successive derivatives of a closed set P_1, if a set P_β does not exist, the previous set $P_{\beta-1}$ consists of a finite set of points, and in case there is a number β such that

$P_\beta \equiv P_{\beta+1}$, the set P_β is perfect. We thus obtain a proof of the theorem already established in § 89, that every p-dimensional set is the sum of a perfect set and an enumerable set. A proof has been* given by de la Vallée Poussin of the second theorem of § 82, applicable to closed sets in any number of dimensions, in which the theorems of § 75 are employed.

112. *Every set of points can be expressed as the sum of a set which is dense in itself, and of an enumerable set which is non-dense in every perfect set.*

In the first place, if a set is non-dense with respect to every perfect set, it is enumerable. For, if a set is not enumerable, there is a set of points of unenumerability which is perfect, and contains points of the set and, in general, others not in the set. The given set would be everywhere dense in this perfect set, contrary to the hypothesis that it is non-dense in every perfect set.

Let E be a given set and consider a point p, whether belonging to E or not, which is such that, in every neighbourhood of p, there is a perfect set with respect to which E is everywhere dense

Let P be the set of all such points p. Any perfect set in which E is everywhere dense must be contained in P; hence P has no isolated points. moreover P is clearly a closed set, and therefore it is perfect. It will be shewn that E is everywhere dense in P. If a perfect set be contained in any interval, or domain, contiguous to P, and E were dense with respect to such perfect set, it would be everywhere dense with respect to some perfect portion of such set, and this portion would therefore belong to P, which is impossible; hence E is non-dense in any perfect set contained in a domain contiguous to P. If E were not everywhere dense in P, the part of P in some interval, or cell, would contain no points of E. Any point of P belonging to this part would have a neighbourhood free from points of E, which is contrary to the definition of P. Hence E is everywhere dense in P. Let

$$E = D(E, P) + E_2:$$

so that E_2 consists of those points of E which do not belong to P. It will be shewn that E_2 is non-dense with respect to every perfect set, and consequently enumerable. If there were a perfect set in which E_2 were not non-dense, it would be everywhere dense in a perfect set R, a portion of the former one. If R have points not belonging to P, E, and therefore E_2, is not everywhere dense in R. If R is contained in P, E_2 has no points in R. Therefore E_2 is non-dense with respect to every perfect set.

In the case of a closed set, the part that is dense in itself is the perfect set which is a part of the closed set; the other part, which is irreducible in every interval, or domain, contiguous to the perfect set, is non-dense in every perfect set.

---

* *Intégrales de Lebesgue*, p. 113.

Denjoy*, to whom the above theorem is due, has indicated a method, in the general case, of obtaining by a successive process the enumerable set that is non-dense with respect to every perfect set.

113. If G be any plane set of points, and, corresponding to each point P, of G, a cell with centre P, and sides all equal to $2h$, be contemplated, the set of all points which belong to one or more such cells, when the cells for all the points P, of G, are considered, is said to be a neighbourhood (h) of the set G. The neighbourhood (h) of the set G is said to be complete, or incomplete, according as the cells are taken to be closed, or open.

Every point of the incomplete neighbourhood (h), denoted by H, of a given set G, is an interior point of H. For any point Q, of the set H, is an interior point of a cell, all the interior points of which belong to H, hence a neighbourhood h' $(< h)$ of Q can be determined so that all its points belong to H. Thus the incomplete neighbourhood (h) of any given set G is an open set.

(*a*) *The complete neighbourhood* (h), *denoted by* $\bar{H}$, *of a closed* (*bounded*) *set* G, *is a closed set.*

For, if $Q_1, Q_2, \ldots$ be a sequence of points of $\bar{H}$ which converges to a point $Q\ (x^{(1)}, x^{(2)})$, and if Q_n be denoted by $(x_n^{(1)}, x_n^{(2)})$, there exists a point $P_n\ (\xi_n^{(1)}, \xi_n^{(2)})$, of G, such that $|x_n^{(1)} - \xi_n^{(1)}|$, $|x_n^{(2)} - \xi_n^{(2)}|$ are both $\leq h$. The set of points $\{P_n\}$ has at least one limiting point $P\ (\bar{\xi}^{(1)}, \bar{\xi}^{(2)})$, which is necessarily a point of G. A sequence of increasing integers $m_1, m_2, m_3, \ldots$ can be so chosen that the sequence $P_{m_1}, P_{m_2}, P_{m_3}, \ldots$ converges to P as its sole limiting point.

Since

$$x_{m_s}^{(1)} - h \leq \xi_{m_s}^{(1)} \leq x_{m_s}^{(1)} + h, \text{ and } x_{m_s}^{(2)} - h \leq \xi_{m_s}^{(2)} \leq x_{m_s}^{(2)} + h,$$

for every value of s, we have

$$x^{(1)} - h \leq \bar{\xi}^{(1)} \leq x^{(1)} + h, \text{ and } x^{(2)} - h \leq \bar{\xi}^{(2)} \leq x^{(2)} + h.$$

It now follows that Q is in the complete neighbourhood of G. Therefore the set $\bar{H}$ is closed.

(*b*) *If* G_1, G_2 *be two* (*bounded*) *closed sets which have no point in common, a neighbourhood of* G_1 *can be determined which contains no points of* G_2.

If d be the distance between the sets G_1, G_2, the complete, or the incomplete, neighbourhood (h), where $h < d/\sqrt{2}$, of G_1, contains no points of G_2.

The following theorem can be deduced:

(*c*) *If the* (*bounded*) *closed set* G *consists entirely of interior points of a*

---

* See his memoir "Sur les nombres dérivés," *Journ. de Math.* (7), vol. I (1915), p. 235.

set H, a neighbourhood (h) of G, complete, or incomplete, can be so determined that it consists entirely of interior points of H.

The points of G have a maximum distance ρ from the point $(0, 0)$. Choose a number ρ' $(> \rho)$, and let H_1 be that part of H that consists of points at a distance $\leq \rho'$ from the point $(0, 0)$.

It is clear that every point of G is an interior point of H_1. Let L be the set of points on the boundary of H_1; L is then a closed set. A neighbourhood (h) of G can be determined which, whether complete or incomplete, contains no points of L. Every point of this neighbourhood (h) is an interior point of H_1, and therefore of H. For any such point Q is in the neighbourhood (h) of some point P, of G. If the neighbourhood (h) of P does not consist entirely of interior points of H_1, it must contain one or more points of L: and this is not the case. Therefore Q is an interior point of H_1[*].

THE CLASSIFICATION OF A FAMILY OF SETS OF POINTS

114. It has been shewn, in § 56, that the set $M\,(O_1, O_2, \ldots O_n, \ldots)$, of points each of which belongs to one at least of the open sets

$$O_1, O_2, \ldots O_n, \ldots,$$

is itself an open set.

Let us consider a set $D\,(O_1, O_2, \ldots O_n, \ldots)$ which, when it exists, consists of all the points common to the open sets of the sequence This set belongs to a class of sets which may be denoted by $O^{(d)}$; so that each set of the class is obtained from some sequence of open sets. A set which can be exhibited as $M\,(O_1{}^{(d)}, O_2{}^{(d)}, \ldots)$ belongs to a class of sets which may be denoted by O^{dm}. In the same way, a set which can be exhibited as $D\,(O_1{}^{(dm)}, O_2{}^{(dm)}, \ldots)$ belongs to a class which may be denoted by O^{dmd}. By continuing this process we can obtain an indefinite number of classes of sets; of these, the first is the class O of open sets, the second is the class of sets $O^{(d)}$, the third is the class of sets $O^{(dm)}$, the fourth the class of sets $O^{(dmd)}$; and so on. The class ω may be defined as the class of sets, each of which consists of the points common to all the sets of a sequence, each member of which belongs to one of the classes $1, 2, 3, \ldots$. The class $\omega + 1$ consists of sets each of which consists of the points belonging to one or more of the sets belonging to a sequence of sets of class ω; the class $\omega + 2$ consists of sets each of which consists of the points common to all the sets of a sequence of sets belonging to class $\omega + 1$.

Proceeding in this manner, we may contemplate the existence of a class corresponding to any ordinal number of the second class.

---

[*] See Hobson, *Proc. Lond. Math. Soc.* (2), vol. xiv (1914), p. 150. The theorem (*c*) was first proved by Bolza in a different manner; see *Vorlesungen über Variationsrechnung*, p. 155.

This classification has been given by Hausdorff[*], who applies the term Borel-set to denote any member of any of these classes, all of which he shews exist. He has proved that every Borel-set is either enumerable (or finite), or has the cardinal number of the continuum. This is a generalization of the theorem that every closed set, which is an $O^{(d)}$, has one of the cardinal numbers a and c. For the whole class $O^{(d)}$, the theorem was established by W. H. Young (see § 97), in connection with the theory of ordinary inner limiting sets. The result obtained by Hausdorff settles the question as to the possible cardinal numbers of sets of points, in the case of sets belonging to a very extensive class.

A detailed treatment of the possible peculiarities of structure of sets of points in plane space, or in space of higher dimensions, is of the highest importance in relation to the application of the principles of the theory of sets of points to abstract geometry. As a development of this character is beyond the scope of the present work, reference is here made to the account given by Schoenflies[†], where references to the literature of the subject will be found. A memoir[‡] by W. H. and G. C. Young, "On the internal structure of a set of points in space of any number of dimensions," may also be here referred to. Information on the subject will also be found in W. H. Young's treatise "On the theory of sets of points." Reference may also be made to G. N. Watson's tract, "Complex integration and Cauchy's theorem," where a proof will be found of Jordan's fundamental theorem, that a plane is divided into two distinct open connex domains, by means of a closed Jordan curve. Another proof[§] has been given by Brouwer.

SETS OF SEQUENCES OF INTEGERS

115. A theory of sets of sequences of integers, of which the formal character is similar to the theory of sets of points in any number of dimensions, has been developed by Baire[||], with a view to application to the Theory of Functions.

A *group* of integers $(\alpha_1, \alpha_2, \ldots \alpha_p)$, of order p, consists of a system of p positive integers arranged in a definite order.

The group $(\alpha_1, \alpha_2, \ldots \alpha_p)$, of order p, is said to be contained in each of the groups (α_1), (α_1, α_2), $(\alpha_1, \alpha_2, \alpha_3)$, ... $(\alpha_1, \alpha_2, \ldots \alpha_{p-1})$ of orders 1, 2, 3, ... $p-1$, respectively.

---

[*] *Math. Annalen*, vol. LXXVII (1916), p. 430.

[†] See his *Entwickelung der Lehre von den Punktmannigfaltigkeiten*, vol. II, chap. IV.

[‡] *Proc. Lond. Math. Soc.* (2), vol XVI (1917), p. 337.

[§] *Math. Annalen*, vol. LXIX (1910), p. 169, and vol. LXXI (1912), p. 314.

[||] *Comptes Rendus*, vol. CXXIX (1899), p. 946.

A *sequence* of integers $(a_1, a_2, \dots a_p, \dots)$ consists of an infinite number of integers, defined in any manner, and arranged in an order similar to the sequence 1, 2, 3, ... This sequence is said to be contained in each of the groups (a_1), (a_1, a_2), ... $(a_1, a_2, \dots a_p)$, ... Let P be a set of such sequences of integers, and let A be any other sequence of integers; then if, for every n, there are sequences in P, other than A itself, which are contained in the same group, of order n, as A itself, that is sequences having their first n integers the same as the first n integers in A, then the sequence A is said to be a *limit* of the set of sequences P. The sequence A may, or may not, itself belong to P.

The set P is said to be *closed*, in case all its limits belong to it. The set is said to be *perfect* when it is closed, and also every sequence in the set is a limit of the set.

A set E, of groups of integers, is said to be *complete* if, when g is any group of order p, belonging to E, the groups of orders $1, 2, 3, \dots p-1$, which contain g, also belong to E.

A complete set E, of groups of integers, is said to be *closed*, if every group g, belonging to E, contains at least one group of higher order than itself, which is also contained in E.

Having given a complete set of groups E, a sequence A may exist such that all the groups containing A belong to E. The set F, of all sequences such as A, is said to be *determined* by the set of groups E. The set F, if it exists, is closed.

Every closed set of groups E determines a closed set of sequences F, and conversely, every closed set of sequences F is determined by a unique closed set of groups E. In case F is perfect, E is also said to be perfect. In order that E may be perfect, it is necessary and sufficient that every group belonging to E should contain at least two groups of one and the same order superior to its own order, and belonging to E.

If P is a set of sequences, then the set P' of those sequences which are limits of the set P is said to be the derived set of P, and may be denoted by P'. The derived set P' is closed.

The successive derivatives $P'', P''', \dots P^{(\omega)}, \dots P^{(\alpha)}$, of finite or transfinite orders, are then defined as in the theory of sets of points. If P is a closed set of sequences, there exists a number α of the first or the second class, such that $P^{(\alpha)} = P^{(\alpha+1)}$. Unless P is an enumerable set, it can be resolved into the sum of an enumerable set and a perfect set.

Let us consider a perfect set of groups E determining a perfect set of sequences F. A set P, of sequences all belonging to F, is said to be *non-dense* in F, or in E, provided that every group of E contains at least one group of E which contains no sequence of P.

A set of sequences P, all belonging to F, is said to be of the *first category*, relative to F, if there exists an enumerable sequence of sets $P_1, P_2, \ldots P_n, \ldots$, each of which is non-dense in F, and such that each sequence of P is part of one at least of the sets $P_1, P_2, \ldots P_n, \ldots$. The set obtained by removing the set P from F is said to be of the *second category* relative to F. The same generic distinction between sets of the first and of the second categories holds, as in the theory of sets of points.

CHAPTER III

THE METRIC PROPERTIES OF SETS OF POINTS

116. In Chapter II an account has been given of those properties of sets of points which are called descriptive, in order to distinguish them from the metric properties which we now proceed to discuss. In the investigation of descriptive properties the notion of the distance between a pair of points x, x', defined arithmetically by $\left\{\sum_1^n (x_r - x_r')^2\right\}^{\frac{1}{2}}$, has been employed, but, although the descriptive properties of a set are invariant for a large class of transformations, the distance between a pair of points x, x' is not invariant in such transformations.

The earliest theory of the metric properties of sets was originated, for the case of linear sets, by Hankel[*], and was further developed by Harnack[†], Stolz[‡], and by Cantor[§], who extended the conception of the content of a set to the case of sets of points in a domain of any number of dimensions. Although this theory of content has now been almost entirely superseded by the later theory of measure, developed by Borel[||] and Lebesgue[¶], we proceed to give an account of it, not only for historical reasons, but also in order to point out the respects in which it has defects which are remedied in the later theory of the measure of sets of points. A very general theory of measure has been given by Carathéodory[**].

THE CONTENT OF A SET OF POINTS

117. Suppose a linear set of points G to be contained in the finite interval (a, b). Let a system of nets be fitted on to the interval (a, b). In this case it will be convenient to take the meshes of the nets to be closed at both ends. In the net of order r, let there be n_r meshes, and of these suppose ν_r contain points of G within them, or at their ends. Let S_{n_r, ν_r} $(\leq b - a)$ denote the sum of the breadths of the ν_r meshes. It is clear that $S_{n_r, \nu_r} \geq S_{n_{r+1}, \nu_{r+1}}$; and thus the sequence of numbers

$$S_{n_1, \nu_1},\ S_{n_2, \nu_2},\ \ldots\ S_{n_r, \nu_r}\ \ldots,$$

---

[*] *Math Annalen*, vol. XX (1882), p. 86.

[†] *Math. Annalen*, vol. XXV (1885), p 241.

[‡] *Math. Annalen*, vol. XXIII (1884), p. 152; see also Pasch, *Math Annalen*, vol. XXX (1887), p. 142.

[§] *Math. Annalen*, vol. XXIII (1884), p. 473.

[||] See *Leçons sur la théorie des fonctions* (1898).

[¶] See the memoir "Intégrale, longueur, aire," *Annali di Mat.* (3), vol. VII (1902), p. 231; also *Leçons sur l'intégration* (1904).

[**] See *Gottinger Nachrichten*, 1914, p. 404.

which are all > 0, has a definite limit Σ, which is such that $0 \leq \Sigma \leq b - a$. The number $S_{n_r, \nu_r} - \Sigma$ may be made as small as we please, by taking r sufficiently large.

It will be shewn that the number Σ is the same whatever closed system of nets is employed. Suppose two systems of nets to give the values Σ_1, Σ_2, of Σ, respectively. We may denote by $S_{n_r^{(1)}, \nu_r^{(1)}}$, $S_{n_r^{(2)}, \nu_r^{(2)}}$ the values of the sum S_{n_r, ν_r} for the two systems of nets, $\{D_n^{(1)}\}$, $\{D_n^{(2)}\}$.

If a net $D_m^{(1)}$ of one system be superimposed on a net $D_{m'}^{(2)}$ of the other system, we obtain a net which may be denoted by $D_{m, m'}$.

If σ denote the sum of the breadths of those meshes of $D_{n_r^{(1)}, n_s^{(2)}}$ that contain points of G, it is clear that $\sigma \leq S_{n_r^{(1)}, \nu_r^{(1)}}$.

Let r be so large that $S_{n_r^{(1)}, \nu_r^{(1)}} - \Sigma_1 < \epsilon$, an arbitrarily chosen positive number; and let s be so large that the greatest of the breadths of the meshes in the nets $D_s^{(2)}$ is $< \eta$, an arbitrarily chosen positive number. Of the $\nu_s^{(2)}$ meshes in the sum $S_{n_s^{(2)}, \nu_s^{(2)}}$ the number that are not also meshes in the sum σ is at most $n_r - 1$; hence $S_{n_s^{(2)}, \nu_s^{(2)}} < \sigma + n_r \eta$. Therefore

$$\Sigma_2 < S_{n_s^{(2)}, \nu_s^{(2)}} < S_{n_r^{(1)}, \nu_r^{(1)}} + n_r \eta < \Sigma_1 + \epsilon + n_r \eta.$$

Now ϵ and η are arbitrarily small, and it therefore follows that $\Sigma_2 \leq \Sigma_1$. It can similarly be proved that $\Sigma_1 \leq \Sigma_2$; and from the two relations it is inferred that $\Sigma_1 = \Sigma_2$. Therefore Σ is a definite number, independent of the particular system of nets employed in defining it.

We have now established the following theorem:

If G be any given set of points in the interval (a, b), there corresponds to G a definite number Σ, which is such that all the points of G are interior points of a definite number of intervals whose sum exceeds Σ by less than an arbitrarily chosen positive number ϵ, the number of the intervals depending on ϵ.

The number Σ is called the *content* of the set G, and the content may have any value in the closed interval $(0, b - a)$.

Those sets of points for which the content is zero are of special importance in the Theory of Functions. A set of zero content is said to be an *unextended*, or a *discrete*, or an *integrable* set of points.

It is clear that the definition, and the proof of existence given above, are applicable in the case of a bounded set of points in space of any number of dimensions, provided intervals are replaced by cells, and the breadth of an interval is replaced by the product of the lengths of the sides of a cell, which is regarded as the content of the cell. A point of the set that is on the boundary of a cell is also in one adjacent cell at least.

118. ***The content of a set of points is the same as the content of its derivative.***

Let Σ' be the content of G', the derivative of a set G; then the points of G' may be included in a finite set of intervals, or cells, the sum of whose contents is less than $\Sigma' + \delta$, where δ is an arbitrarily chosen positive number. There can be only a finite number of points of G which do not fall within the intervals, or within the cells, that include the points of G'; and this finite number of points may be included in intervals, or cells, the sum of the contents of which is arbitrarily small, say ϵ. All the points of G are now included in a finite number of intervals, or cells, of total content less than $\Sigma' + \delta + \epsilon$; and a series of diminishing values may be assigned to δ and ϵ; each sequence having the limit zero. Therefore both $\Sigma' + \delta + \epsilon$ and $\Sigma' + \delta$ converge to the value Σ', which proves the theorem.

It follows, from this theorem, that the content of any set is the same as that of any of its successive derivatives. In the case of a set which is of the first species, one of the derivatives contains only a finite number of points, and consequently the set must be of zero content.

119. A definition of the content of a linear set of points has been given by Cantor* which, though differing in form from that of Hankel and Harnack, is in reality equivalent to it. Instead of enclosing the points of the set G in a finite number of intervals, Cantor encloses each point of G in an interval 2ρ, of which the point is the middle point, the number ρ being the same for each point of the set, those parts of intervals 2ρ which do not lie within (a, b) being disregarded. We have in this manner obtained an infinite number of overlapping intervals which contain all the points of G, and, as is clear, all the points of G', which is a closed set. If we replace this set of intervals by the set of non-overlapping intervals with the same interior points, each interval of this latter set is $\geqq \rho$. The set, which is non-overlapping, and equivalent to the infinite set, is consequently a finite set, the sum of whose lengths may be denoted by $\Pi(\rho, G)$. When ρ is diminished indefinitely, the number $\Pi(\rho, G)$, which cannot increase as ρ is diminished, must have a definite lower limit, which defines the content of either of the sets G and G'. Since the infinite set of intervals which has been employed only covers a finite number of detached lengths, this definition is equivalent to that of Hankel and Harnack. Cantor applied this definition to the case of a set of points in a p-dimensional continuum, by enclosing each point in a "sphere" of radius ρ, with its centre at the point; the content is then the lower limit of the volume made up of the points contained within the spheres.

The essential point in the above definition of the content of a set of points is that all the points are contained in a finite number of intervals, or cells, which therefore contain all the limiting points†; and the lower limit of the sum of the contents of these intervals, or cells, is taken as defining the content of the set. If the points are contained, from the commencement, in an infinite number of intervals, or cells, which are of unequal span, in accordance with some prescribed law, and the spans of these intervals, or cells, are then diminished, each one in a prescribed manner tending to the limit zero, then the limit of the sum of those parts of the fundamental interval, or cell, which are included in the infinite set of intervals, or cells, is not necessarily equal to the content, as above defined. For example, let us consider the set of rational points in the interval (0, 1). These points can be arranged in enumerable order P_1, P_2, P_3, ...: now enclose P_1 in an interval of length $\epsilon/2$, P_2 in an interval $\epsilon/2^2$, &c., P_n in an interval of length $\epsilon/2^n$, and so on; the total content of these intervals cannot exceed $\epsilon\Sigma 1/2^n$, or ϵ; and this has the limit zero, as ϵ is diminished towards zero. On the other hand, the content of the set of rational points is the same as that of the derived set; but this consists of all the points of the interval (0, 1), and is therefore unity. In general, any enumerable set of points can be contained in an infinite number of intervals, or cells, which have a total content that is arbitrarily small, and has the limit zero; whereas the content of the set is not in general zero.

---

* *Math. Annalen*, vol. XXIII (1884), p. 473.

† See Harnack, *Math. Annalen*, vol. XXIII (1884), p. 241.

A completely satisfactory definition of the content of a set of points of the most general character should satisfy the condition of affording a consistent generalization of the notion of the length of a continuous linear set of points, or of the notions of area and volume, in the case of sets of points in two, or in three, dimensions. In the case of closed sets, the definition given above leaves nothing to be desired in this respect; but in the case of unclosed sets, the definition leads to consequences which are at variance with the fundamental properties of lengths, areas, and volumes, as understood for the case of continuous domains. If G_1, G_2 are two complementary sets of points in the continuous interval (0, 1), then, in order that the contents of the sets G_1, G_2 may accord with a generalization of the notion of length, their sum should be unity; however, when G_1 and G_2 are unclosed, this condition is in general not satisfied by the definition given above. For example, if G_1 consists of the rational points, and G_2 of the irrational points, each of the two sets G_1, G_2 has its content unity, the same as that of the interval (0, 1) itself. Again, let us consider an everywhere dense set of non-overlapping intervals contained in (0, 1); then the internal points of these intervals form an open set G_1, of which the derivative consists of all the points of the interval (0, 1); the external and the end-points of the intervals forming a non-dense closed set G_2. It will be shewn subsequently that the everywhere dense set of non-overlapping intervals can be so chosen that the limit of the sum of their lengths is an arbitrary number l, where l is subject to the condition $0 < l \leq 1$; whereas the content of the set G_1 is, in accordance with the definition given above, always unity, and therefore may differ from the sum of the contents of the sets of points contained in the separate intervals. To obtain the content of the closed set G_2, cut off, from each of the intervals which define G_1, the $1/2n$th part of its length at each end; the limiting sum of the intervals, so restricted, is $l(1-1/n)$. Of these restricted intervals, a finite number can be so taken that their sum is $> l(1-1/n)-\epsilon$, and $< l(1-1/n)$, where ϵ is an arbitrarily chosen positive number. All the points of G_2 are now enclosed in the finite set of intervals which is complementary to the finite set of restricted intervals. The sum of these complementary intervals $< 1-l(1-1/n)+\epsilon$, and $> 1-l(1-1/n)$; the sum has for its lower limit the number $1-l$, which is therefore the content of G_2. The sum of the contents of G_1, G_2 is therefore not equal to unity, which is nevertheless the content of

$$G_1 + G_2 \equiv (0, 1).$$

THE PROBLEM OF MEASURE

120. The problem of assigning definite numerical measures to sets of points defined in linear, plane, solid, or p-dimensional. space is taken to require that the measure of a set shall satisfy the following conditions:

(1). The measure of a set shall be in accord with the usual notions of length, area, volume, of an interval, a rectangle, or a cell in three or more dimensions.

Accordingly there must exist separate systems of measures for linear, plane, or p-dimensional, sets of points, corresponding to generalizations of the measures of length, area, or volume.

(2). The linear measure of the set of points in a linear interval (a, b) is taken to be $b-a$, whether the set includes neither, one. or both, of the end-points a and b, of the interval.

The p-dimensional measure of the set of points in the cell

$$(a^{(1)}, a^{(2)}, \dots a^{(p)};\ b^{(1)}, b^{(2)}, \dots b^{(p)})$$

is taken to be the product

$$(b^{(1)}-a^{(1)})\,(b^{(2)}-a^{(2)})\dots(b^{(p)}-a^{(p)}),$$

whether the set of points includes none, some, or all, of the points on the boundary of the cell.

Thus, when $p=2$, the measure of the set of points in the rectangle is in agreement with the ordinary measure of its area; when $p=3$, the measure of the set of points in the rectangular parallelepiped is in agreement with the ordinary measure of its volume.

It will be observed that, in accordance with this postulation, the p-dimensional measure of a q-dimensional cell, when $q<p$, is zero. Thus a linear interval has the plane measure zero; and a plane rectangle has the three-dimensional measure zero.

(3). The measure of a set of points is to be a number dependent on the set, such that the measure of the sum of two sets, which have no point in common, is the sum of the measures of the two sets. It then follows that the measure of the sum of any finite number of sets, no two of which have a point in common, is the sum of the measures of the sets. The measure of a set being regarded as a function of the set, is thus required to be an *additive* function, *i.e.* a function such that its value for the set $E_1 + E_2$ is the sum of its values for E_1 and E_2.

It follows that, if a set F be a component of a set E, the measure of the set $E - F$ is to be the excess of the measure of E over that of F. In case E and F have the same measures, the set $E - F$ is to have the measure zero.

In particular, the p-dimensional measure of a $(p - 1)$-dimensional set is zero. For such a set may be taken to be on the boundary of a p-dimensional cell; and, in accordance with (1), the measure of the set of points interior to the p-dimensional cell is unaltered by addition of the $(p - 1)$-dimensional set on its boundary; consequently the p-dimensional measure of this last set must be zero.

(4). The measure of the sum of an enumerably infinite sequence of sets, no two of which have a point in common, is to be the limiting sum of the measures of the sets, whenever that limiting sum exists. This may be expressed as the postulation that the measure of a set shall be a *completely additive* function of the set.

(5). The measures of two sets such as are obtained from one another by a congruent transformation are to be identical.

Congruent sets are such that, to each pair of points P, Q, of one of the sets there corresponds a unique pair of points P', Q', of the other set, such that the distance of P' from Q' is the same as that of P from Q.

In order that a system of measurement of sets may be set up, for each number of dimensions, which shall satisfy the above postulates, definitions will be introduced which will apply to a certain category of sets, called measurable sets, and which, for sets of this category, will be shewn to provide a function of the set that satisfies the above postulates. It will appear that the class of measurable sets, in linear, plane, or higher dimensional, space includes all the sets which are defined by certain prescribed methods. The possibility of extending the theory to cases of non-measurable sets will be left out of account*.

---

* A definition of a non-measurable set has been given by Van Vleck, *Trans. Amer. Math. Soc.* vol. IX (1908), p. 237, but this definition requires an infinite number of arbitrary acts of choice. Such sets have also been defined by Lebesgue, *Bull. de la soc. math. de France*, vol. XXXV 907), p. 209.

THE MEASURES OF OPEN AND CLOSED SETS

121. A linear set of points O, open relatively to a linear interval (a, b) which contains the set, is the sum of a unique enumerable set of open intervals.

In accordance with (3), the measure of O must be taken to be the limiting sum of the measures of the intervals of this set. That this limiting sum is finite, and cannot exceed l, the measure of (a, b), is seen by considering the first n of the intervals, as arranged in order: the sum of these n intervals is less than l by the sum of the lengths of the finite set of intervals complementary to them, and this holds good however large n may be.

Every open interval may be regarded as consisting of the points of an enumerable set of closed intervals, no two of which overlap one another, although they may have an end-point in common; therefore any open set may be regarded as consisting of the points of an enumerable set of non-overlapping closed intervals. The measure of the open set can then be regarded as the limiting sum of the measures of such a system of non-overlapping closed intervals.

In accordance with (1), it makes no difference that pairs of the intervals may have an end-point in common.

If G be any closed linear set in (a, b), the complementary set $C(G)$ is open relatively to (a, b), and in accordance with (2), the measure of G must be given by $m(G) = l - m\{C(G)\}$; where the measures of G and $C(G)$ are denoted by $m(G)$ and $m\{C(G)\}$ respectively.

122. It has been shewn in § 109 that, in space of any number of dimensions, an open set O, in a cell (a, b), is constituted by the points of an enumerable set of closed cells. In accordance with (3), the measure of O must be the limiting sum of the measures of these closed cells. That this limiting sum is finite, and does not exceed the measure l, of the cell (a, b), is proved as in the case of a linear set.

If G be a closed set, in the cell (a, b), the measure of G must, in accordance with (2), be given by $m(G) = l - m\{C(G)\}$; the set $C(G)$ being open relatively to (a, b).

Since the set of closed cells, or intervals, which constitutes a given open set O is not unique, it must be shewn that the measure of O, defined as the limiting sum of the measures of the closed cells, or intervals, is independent of the particular sets of such cells, or intervals. In order to shew this, the following theorem, somewhat more general than is required for this particular purpose, will be proved:

If an open set O consist of all the points of an enumerable set $\{\Delta_n\}$, of *non-overlapping closed cells, and if* $\{\Delta_n'\}$ *be any other set of non-overlapping closed cells such that every point of* O *belongs to at least one of the closed cells* $\{\Delta_n'\}$, *the limiting sum of the measures of* $\{\Delta_n'\}$ *cannot be less than that of the measures of* $\{\Delta_n\}$.

Assume, if possible, that the limiting sum of the measures of the cells of $\{\Delta_n\}$ exceeds that of the measures of $\{\Delta_n'\}$. A finite set $\Delta_1, \Delta_2, \dots \Delta_m$, of the cells $\{\Delta_n\}$, may then be so determined that $\sum_{r=1}^{r=m} m(\Delta_r)$ is greater than the limiting sum of the measures of the cells $\{\Delta_n'\}$.

A cell $\bar{\Delta}_n'$ containing Δ_n' in its interior, and with the same centre, can be defined such that $m(\Delta_n') - m(\Delta_n') = \epsilon/2^n$; where ϵ is any chosen positive number; and this can be done for each of the cells $\{\Delta_n'\}$. The limiting sum of the measures of the cells $\{\bar{\Delta}_n'\}$ exceeds that for the cells $\{\Delta_n'\}$ by ϵ. If ϵ be sufficiently small, $\sum_{r=1}^{r=m} m(\Delta_r)$ is greater than the limiting sum of the measures of the cells $\{\bar{\Delta}_n'\}$. The set of points in $\Delta_1, \Delta_2, \dots \Delta_m$ is a closed set G_m; and since each point of G_m is interior to a cell of $\{\bar{\Delta}_n'\}$, from the Heine-Borel theorem (§ 74) it follows that a finite set of the cells $\{\bar{\Delta}_n'\}$ can be determined so as to contain all the points of G_m. The total measure of the cells of this finite set cannot be less than $\sum_{r=1}^{r=m} m(\Delta_r)$; and this is contrary to the supposition made above.

It follows from this theorem that, if $\{\Delta_n\}$, $\{\Delta_n'\}$ be two sets of closed non-overlapping cells, each of which constitutes a given open set O, the limiting sums of the measures of the cells of each set must be the same; for, by the theorem, it is impossible that either of these should be less than the other.

The unique measure of an open bounded set O *is thus defined** *as the limiting sum of the measures of the closed cells, or intervals, of a set, all the points of which constitute the set* O.

The measure of a bounded closed set G *is then defined to be the excess of the measure of the fundamental cell, or interval, in which* G *is contained, over the measure of the open set* $C(G)$, *which is the complement of* G *with regard to the fundamental cell, or interval.*

All the points of the closed set G are in the finite set of cells, or intervals, complementary to the cells, or intervals, $\Delta_1, \Delta_2, \dots \Delta_n$. Thus $m(G)$ is the lower limit of the sum of the measures of this finite set, as n is indefinitely increased. It is thus seen that the measure of a closed set as here defined is identical with the content of such a set as defined in § 117. It has been seen that *any open* set O contains a closed set whose measure is arbitrarily little less than $m(O)$.

---

* de la Vallée Poussin, *Intégrales de Lebesgue*, p. 22.

123. It follows, as a particular case of the theorem proved in § 122, that, if O_2 be an open set which contains another open set O_1,

$$m(O_2) \geqq m(O_1).$$

For, if O_1 is constituted by a set of closed cells $\{\Delta_n\}$, and O_2 by a set of closed cells $\{\Delta_n'\}$, the limiting sum of the measures of the cells Δ_n' cannot be less than that of the measures of the cells Δ_n; and thus $m(O_2)$ cannot be less than $m(O_1)$.

If $O_1, O_2, \dots O_n, \dots$ be a sequence of open sets, all contained in a finite cell, or interval, then

$$m(O_1) + m(O_2) + \dots + m(O_n) + \dots \geqq m[M(O_1, O_2, \dots O_n, \dots)].$$

It has been shewn in § 56 that $M(O_1 . O_2, \dots O_n; \dots)$ is an open set; let $\{\Delta_m\}$ be a sequence of closed cells which constitute it. Let O_n be constituted by a set of closed cells $\{\Delta_{nm}\}$, where $m = 1, 2, 3, \dots$.

The set of cells $\{\Delta_{nm}\}$, where n and m have all integral values, is an enumerable set of cells, the limiting sum of whose measures is

$$m(O_1) + m(O_2) + \dots + m(O_n) + \dots.$$

By the theorem proved in § 122, this cannot be less than the limiting sum of the measures of $\Delta_1, \Delta_2, \dots \Delta_m, \dots$; or than the measure of

$$M(O_1, O_2, \dots O_n, \dots).$$

It has here been assumed that if $\{c_{nm}\}$ is a double sequence of positive numbers, $\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{nm}$ is unaltered by changing the order of the terms.

If $O_1, O_2, \dots O_n, \dots$ be a sequence of open sets, all contained in a finite cell, or interval, and such that no two of them have a point in common, then

$$m(O_1) + m(O_2) + \dots + m(O_n) + \dots = m[M(O_1, O_2, \dots O_n, \dots)].$$

In this case, no two of the cells Δ_{nm} overlap one another, and the set

$$M(O_1, O_2, \dots O_n, \dots)$$

is constituted by the enumerable set of cells Δ_{nm}, where

$$n = 1, 2, 3, \dots; \quad m = 1, 2, 3, \dots.$$

Both $\sum_{n=1}^{\infty} m(O_n)$ and $m[M(O_1, O_2, \dots)]$ are equal to the limiting sum of the measures of the cells Δ_{nm}.

It has thus been shewn that, for open sets, the measure is a completely additive function; thus satisfying the postulates (3) and (4), of § 120.

THE CONTENT OR MEASURE OF A CLOSED SET

124. In a linear interval, of length l, there can be defined a set of non-overlapping intervals, everywhere dense in the interval, and such that the sum of the lengths of the intervals is less than an arbitrarily chosen number ϵ. To establish this apparently paradoxical theorem, let us consider an enumerable set of points $x_1, x_2, \ldots x_n, \ldots$ everywhere dense in the interval; for example, the rational points of the interval. With each point x_n as centre, an interval of length $\epsilon/2^n$ may be defined; any part of the interval which is not in the fundamental interval may be omitted. The overlapping set of open intervals obtained by taking all the values of n is equivalent to a non-overlapping set of open intervals Δ, of total measure less than ϵ, which is everywhere dense in the fundamental interval. The points of that interval complementary to the set Δ form a non-dense closed set of content, or measure, greater than $l - \epsilon$. This can be made arbitrarily near to l, by taking ϵ small enough.

The following general theorem will be established:

The content of a non-dense linear closed set is zero, in case the set is enumerable; and, in case the set is unenumerable, its content may be zero, or may have any value less than the length of the interval in which the set is contained.

The content, or measure, of a closed set is the sum of the measures of its perfect component and of its enumerable component; and this last is zero. If the set is enumerable, it has no perfect component, and therefore its content is zero.

Let $\delta_1, \delta_2, \delta_3, \ldots$ denote the lengths of the intervals contiguous to a non-dense closed set G, in a fundamental interval of length l.

Let $\delta_1 = \lambda_1 l,\ \delta_2 = \lambda_2 (l - \delta_1),\ \delta_3 = \lambda_3 (l - \delta_1 - \delta_2), \ldots$

and generally, $\delta_n = \lambda_n (l - \delta_1 - \delta_2 - \ldots - \delta_{n-1})$;

where the numbers $\lambda_1, \lambda_2, \ldots \lambda_n, \ldots$ are all between 0 and 1.

We find at once

$$\delta_n = \lambda_n (1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_{n-1})\, l$$

and hence

$$l - (\delta_1 + \delta_2 + \ldots + \delta_n) = (1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_n)\, l.$$

The content of the set G is therefore* l multiplied by the limit of the product $(1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_n)$.

The values of $\lambda_1, \lambda_2, \ldots \lambda_n, \ldots$ can be so chosen that the content of the set is zero; for example if $\lambda_1 = \lambda_2 = \lambda_3 = \ldots$.

If $\lambda_1 = \theta^2, \lambda_2 = \theta^2/2^2, \ldots \lambda_n = \theta^2/n^2, \ldots$; where $0 < \theta < 1$, the content of the set is $l \sin(\pi\theta)/\pi\theta$. By choosing a sufficiently small value of θ, this may be made as near as we please to l. By proper choice of θ, the content may have any value less than l.

125. *Any linear closed set G that has interior points contains a non-dense closed set H, such that $m(G) - m(H)$ is less than an arbitrarily chosen number ϵ.*

The set G may contain a finite, or an enumerable, set of closed intervals,

---

* See Harnack, *Math. Annalen*, vol. XIX (1882), p. 239.

$\delta_1, \delta_2, \ldots \delta_n, \ldots$. Remove from G all the interior points of δ_n except those belonging to a non-dense closed set g_n, of content greater than $m(\delta_n) - \epsilon/2^n$. The set which consists of G, with the interior points of all the intervals $\{\delta_n\}$ removed, except the points of $g_1, g_2, \ldots g_n, \ldots$, is a non-dense closed set H, such that $m(G) - m(H) < \epsilon$.

Similar results hold for plane closed sets, or for closed sets in space of any number of dimensions. It will be sufficient to consider the case of plane closed sets.

In a cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ an everywhere dense set of cells can be defined, of which the total measure is either that of the given cell, or has any less value > 0.

In the linear interval $(a^{(1)}, b^{(1)})$ an everywhere dense set of intervals $\{\delta_n\}$ can be defined, as in § 124, of which the total measure has a prescribed value > 0, and not exceeding $b^{(1)} - a^{(1)}$. Similarly a set of intervals $\{\delta_n'\}$ may be defined in the linear interval $(a^{(2)}, b^{(2)})$, of total measure > 0, and not exceeding $b^{(2)} - a^{(2)}$.

Consider a cell for which the projections of the sides on the two axes of coordinates are $\delta_n, \delta'_{n'}$; where n and n' are any pair of integers. The set of all such cells is everywhere dense in the fundamental cell, and the sum of their measures has a value > 0, and not exceeding the measure of the fundamental cell. The set of cells may clearly be chosen so that the total measure has any such prescribed value. The set of points which is complementary to the set of open cells so defined is a closed non-dense set, of which the measure may be zero, or may have any prescribed value less than the measure of the fundamental cell.

A closed plane set G, that has interior points, contains a non-dense closed component H, such that $m(G) - m(H)$ is less than an arbitrarily prescribed number ϵ.

It has been shewn, in § 106, that such a set as G contains a finite, or an enumerable, set of closed connex domains $\{D_n\}$. The interior points of D_n are constituted by an enumerable set of closed cells. Hence the interior points of all the domains $\{D_n\}$ together consist of an enumerable set of closed cells $\{\Delta_n\}$. We can remove from Δ_n all its points except those of a non-dense closed set of measure $> m(\Delta_n) - \epsilon/2^n$. When this has been done for all the cells $\{\Delta_n\}$, we have remaining a part H, of G, such that $m(G) - m(H)$ is less than ϵ; for the sets of points removed have a total measure $< \Sigma\epsilon/2^n$, or ϵ. The set H is non-dense in the fundamental cell.

EXAMPLES

1. The perfect set of points defined by $x = \frac{c_1}{3} + \frac{c_2}{3^2} + \ldots + \frac{c_n}{3^n} + \ldots$, where the numbers $c_1, c_2, \ldots$ have each one of the values 0, 2 (see Ex. 1, § 83), has the content zero. For the limit of the sum of the lengths of the complementary intervals is unity.

2. The non-dense closed set considered in Ex. 3, § 83, has the content zero. For, after k operations, the sum of the exempted segments is

$$\frac{1}{m}+\frac{m-1}{m^2}+\frac{(m-1)^2}{m^3}+\ldots+\frac{(m-1)^{k-1}}{m^k}, \text{ or } 1-\left(\frac{m-1}{m}\right)^k.$$

When k is increased indefinitely, the limit of the sum of the free intervals is 1.

3. The non-dense closed set considered in Ex. 4, § 83, has a content between 0 and 1. After k operations, the sum of the exempted segments is

$$\frac{1}{m}+\frac{m-1}{m^3}+\frac{(m-1)(m^2-1)}{m^6}+\ldots+\frac{(m-1)(m^2-1)\ldots(m^{k-1}-1)}{m^{\frac{1}{2}k(k+1)}},$$

or

$$1-\left(1-\frac{1}{m}\right)\left(1-\frac{1}{m^2}\right)\ldots\left(1-\frac{1}{m^k}\right).$$

The limit of the sum of the exempted intervals is $1-\prod_{k=1}^{\infty}\left(1-\frac{1}{m^k}\right)$, and therefore the content of the set of points is $\prod_{k=1}^{\infty}\left(1-\frac{1}{m^k}\right)$, which is between 0 and 1, depending upon the value of m. By taking m sufficiently great, the content of the set may be made arbitrarily near to unity.

THE EXTERIOR AND INTERIOR MEASURES OF A SET

126. Let G be any linear set of points in a given interval (a, b), and let a finite, or an infinite, set of non-overlapping open intervals be defined, such that every point of G is in one of these intervals. The point a is regarded as an interior point of any interval of which it is an end-point; and a similar remark applies to b. The set of intervals constitutes an open set which contains G. The sum, or limiting sum, of the lengths of the intervals has a definite value, not greater than l. The lower boundary of this sum or limiting sum, for all possible such sets of intervals, is a number which is called the *exterior measure of the set* G; and this may be denoted by $m_e(G)$.

Alternatively, the exterior measure $m_e(G)$, of the set G, may be defined as the lower boundary of the sum, or limiting sum, of the lengths of the intervals of a set of non-overlapping closed intervals such that every point of P is in a closed interval of the set, being either an interior point, or an end-point, when all such systems of closed intervals are taken into account.

To shew the equivalence of the two definitions, let L_1, L_2 denote the lower boundaries defined in accordance with the first, and the second, definition respectively. A set of non-overlapping open intervals containing G can be defined, such that the sum, or limiting sum, of their lengths is $< L_1+\epsilon$; where ϵ is an arbitrarily chosen positive number. Since each open interval is constituted by an enumerable set of non-overlapping closed intervals, the set of open intervals can be replaced by a set of closed intervals which contain G; and the limiting sum of the lengths of the intervals is the same in the two cases. This shews that $L_2 \leq L_1$, since we have a set of closed intervals satisfying the requisite condition, the limiting sum of whose lengths is $< L_1 + \epsilon$. Again, it is impossible that $L_2 < L_1$; for if a set of non-overlapping closed intervals containing all the points of G existed, the limiting sum of whose lengths is $< L_1$, we could enclose each of these intervals in an open interval, of length exceeding that of the closed interval by less than an arbitrarily chosen number, and in this way we could obtain an overlapping set of open intervals containing G, and such that the limiting sum of their lengths would be $< L_1$. This overlapping set might be replaced by the equivalent set of non-overlapping intervals, and the limiting sum of the latter intervals cannot exceed that of the former, as has been shewn in the first theorem of § 123. We should therefore have a set of non-overlapping open intervals, enclosing G, the limiting sum of whose lengths would be $< L_1$, contrary to the fact that L_1 is the lower boundary of all such limiting sums.

Let $C(G)$ be the complementary set of G, relatively to the interval (a, b) of length l, in which G is contained. If $m_e\{C(G)\}$ denote the exterior measure of $C(G)$, the number $l - m_e\{C(G)\}$ is taken to define the *interior measure* of the set G; and this interior measure may be denoted by $m_i(G)$.

127. In the case of a set G contained in a finite cell of any number of dimensions, a connex open domain takes the place of an open interval; and thus the exterior measure $m_e(G)$ is defined as the lower boundary of the limiting sums of the measures of connex open domains which form a finite, or an infinite, sequence containing all the points of G, when all such sets of connex open domains are taken into account. Again, since any open set is constituted by a sequence of closed cells, the exterior measure $m_e(G)$ may also be defined as the lower boundary of the limiting sums of the measures of sequences of closed cells, each sequence containing all the points of G in the interior and on the boundaries of the cells, when all such sequences are taken into account That this second form of the definition is equivalent to the first may be shewn in exactly the same manner as in the case of a linear set; for every sequence of closed cells forms part of a sequence of open cells, such that the difference of the sums of the measures of the cells in the latter sequence and in the former is arbitrarily small.

As before, the interior measure $m_i(G)$ of G is defined by

$$m_i(G) \equiv l - m_e\{C(G)\};$$

where l is the measure of the cell in which G is contained.

The above definitions of the exterior and interior measures of a bounded set of points in any number of dimensions are equivalent to the following statements:

The exterior[*] measure $m_e(G)$ of a bounded set G is the lower boundary

---

[*] **This form of the definition is due to de la Vallée Poussin, see *Intégrales de Lebesgue*, p. 22.**

of the measures of open sets which contain G. The interior measure $m_i(G)$ of a bounded set G is the upper boundary of the measures of closed sets contained in G.*

For, as the exterior measure $m_e\{C(G)\}$ is the lower boundary of open sets which contain $C(G)$, it follows that $l - m_e\{C(G)\}$ is the upper boundary of the measures of the closed sets complementary to these open sets.

If a set G_1 contains a set G_2, it is clear from the definition that

$$m_e(G_1) \geqq m_e(G_2).$$

MEASURABLE SETS OF POINTS

128. It appears from the above definitions that every bounded set of points has definite exterior and interior measures.

When the exterior and interior measures of a set G, of points in p dimensions, are equal to one another, the set G is said to be measurable, and the number $m_e(G) \equiv m_i(G)$ is defined to be the measure of G. When G is measurable its measure is denoted by $m(G)$.

It is clear from this definition that, if G is measurable, so also is $C(G)$.

It will be shewn that this definition satisfies the conditions stated in § 120, whenever it is applicable. If α is a set of non-overlapping cells, or intervals, enclosing a set of points G, measurable, or not, and β is a similar set for $C(G)$, then all the points of the fundamental cell, or interval, are enclosed in the set made up of α and β. It follows that $m(\alpha) + m(\beta) \geq l$; and since $m_e(G)$, $m_e\{C(G)\}$ are the lower boundaries of $m(\alpha)$, $m(\beta)$, respectively, we have $m_e(G) + m_e\{C(G)\} \geqq l$, and therefore $m_e(G) \geqq m_i(G)$.

The condition that a set G is measurable, in the sense above defined, may be stated in either of the following forms:

A set of points G is measurable if an open set O, containing G, and a closed set H, contained in G, can be so determined that $m(O) - m(H)$ is less than an arbitrarily prescribed positive number ϵ.

A set of points G is measurable if its points can be enclosed in a finite, or an infinite, set α, of open intervals (in the case of linear sets), or of open connex domains (in the case of sets in space of higher dimensions), and if $C(G)$ can be similarly contained in a set β, such that the sum, or limiting sum, of the measures of the open intervals, or connex domains, which contain all those points which are common to α and β is arbitrarily small.

The set of points $D(\alpha, \beta)$ forms an open set (§ 56) which, if the condition be satisfied, may be taken to have its measure $< \epsilon$. As is easily seen by considering the three sets α, β, $D(\alpha, \beta)$, we have

$$m(\alpha) + m(\beta) = m\{D(\alpha, \beta)\} + l;$$

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. II (1905), p. 28.

where l is the measure of the cell, or interval, in which the given set is contained. It then follows that

$$\{m(\alpha) - m_i(G)\} + \{m(\beta) - m_e[C(G)]\} < \epsilon.$$

Since $m(\alpha) - m_i(G)$, $m(\beta) - m_e[C(G)]$ are not negative, it follows that each of them is $< \epsilon$; also $m(\alpha) \geqq m_e(G) \geqq m_i(G)$, hence $m_e(G) - m_i(G) < \epsilon$. Since this holds for every value of ϵ, we have $m_e(G) = m_i(G)$; and thus the condition stated is sufficient that the set G may be measurable. If G be assumed to be measurable, α and β can be so chosen that

$$m(\alpha) - m(G) < \tfrac{1}{2}\epsilon,$$

and
$$m(\beta) - m[C(G)] < \tfrac{1}{2}\epsilon;$$

we then have $m(\alpha) + m(\beta) < \epsilon + l$, which is equivalent to the condition that $m\{D(\alpha, \beta)\} < \epsilon$; and thus the condition is necessary.

129. It must be shewn that the definitions, given in §§ 120 and 122, of the measures of open, closed, or partly closed, cells or intervals, and of open and closed sets, are consistent with the definition of a measurable set given in § 128, that is, that these sets have their exterior and interior measures equal to one another.

The cell $(a^{(1)}, a^{(2)}, \dots a^{(p)} : b^{(1)}, b^{(2)}, \dots b^{(p)})$ is contained in the cell

$$\left[a^{(1)} - \frac{1}{2n}(b^{(1)} - a^{(1)}),\ a^{(2)} - \frac{1}{2n}(b^{(2)} - a^{(2)}), \dots : \right.$$
$$\left. b^{(1)} + \frac{1}{2n}(b^{(1)} - a^{(1)}),\ b^{(2)} - \frac{1}{2n}(b^{(2)} - a^{(2)}), \dots\right]$$

and contains the cell obtained by writing $-n$ instead of n.

The measures of these cells are $\left(1 + \frac{1}{n}\right)^p l$, $\left(1 - \frac{1}{n}\right)^p l$, respectively, and thus differ from one another by $< \epsilon$, if n be chosen sufficiently large. Therefore the exterior and interior measures of a cell are identical with l, the measure of the cell, as defined in § 120.

It is clear that the exterior measure of an open set O is identical with the measure of the set, as defined in § 121, for it has been shewn that no open set of measure less than that of the set can contain O. Also O contains the closed set formed by a finite number of the closed cells, or intervals, which constitute O, and the sum of the measures of these cells, or intervals, converges to the measure of O; thus O is contained in an open set, consisting of O itself, and contains a closed set, such that the difference of the measures of the open and closed sets is arbitrarily small; therefore both the exterior and interior measures of O are identical with the measure of O, as defined in § 121. Since a closed set is the complement of an open set O, it follows that the closed set is measurable. It has thus been proved that:

Every bounded open set, and every bounded closed set, is measurable.

It will now be shewn that:

Every enumerable set of points is measurable, and its measure is zero.

Let $P_1, P_2, \dots P_r, \dots$ denote the points of the set. Each point P_r can be enclosed in an open cell, or interval, with P_r as centre, and of measure $\epsilon/2^r$. In accordance with a theorem proved in § 123, the open set of points consisting of all the points belonging to one or more of these cells, or intervals, has its measure $\leq \sum_{r=1}^{\infty} \epsilon/2^r$, or $\leq \epsilon$. It follows that the exterior measure of the set is zero, and thus that the interior measure is also zero, and that the set is measurable, with measure equal to zero.

130. *If $G_1, G_2, \dots G_n, \dots$ is an enumerable (or a finite) sequence of measurable sets, all contained in a finite cell, or interval, the set*

$$M(G_1, G_2, \dots G_n, \dots),$$

which consists of all points that belong to one or more of the sets, is measurable.

Let G_1 and $C(G_1)$ be contained in open connex domains, or open intervals, of sets α_1, β_1, of such domains, or intervals, so defined that the measure of $D(\alpha_1, \beta_1)$ is $< \frac{1}{2}\epsilon$. Let α_2, β_2 be similar sets of connex open domains, or intervals, which contain G_2 and $C(G_2)$ respectively, and are such that $m[D(\alpha_2, \beta_2)] < \epsilon/2^2$. Let $\alpha_2' = D(\alpha_2, \beta_1)$, and $\beta_2' = D(\beta_2, \beta_1)$. For G_3 and $C(G_3)$ we similarly consider sets α_3, β_3, of open domains, or intervals, such that $m[D(\alpha_3, \beta_3)] < \epsilon/2^3$; and let

$$\alpha_3' = D(\alpha_3, \beta_2'), \qquad \beta_3' = D(\beta_3, \beta_2');$$

and so on, for all the sets $G_4, G_5, \dots$. The points of $M(G_1, G_2, G_3, \dots)$ are all contained in the sets of domains, or intervals, $\alpha_1, \alpha_2', \alpha_3', \dots$; and $C\{M(G_1, G_2, \dots)\}$ is contained in β_i', whatever value i may have.

The two sets of domains, or intervals, which enclose $M(G_1, G_2, \dots)$ and its complement, have in common a set of points of which the measure is less than

$$\epsilon/2 + \epsilon/2^2 + \dots + \epsilon/2^i + m(\alpha'_{i+1}) + m(\alpha'_{i+2}) + \dots.$$

The series $\Sigma m(\alpha')$ is convergent, since each term is positive, and the sum of any number of terms is less than a fixed finite number. Therefore the number i can be so chosen that $m(\alpha'_{i+1}) + m(\alpha'_{i+2}) + \dots$ is less than ϵ.

Now $M(G_1, G_2, \dots)$ and its complement have been enclosed in sets of intervals, or cells, such that the set common to both sets has measure $< 2\epsilon$. Since ϵ is arbitrary, it has been proved that $M(G_1, G_2, \dots)$ satisfies the criterion that it is measurable.

If $G_1, G_2, \dots G_n, \dots$ be measurable sets in linear or higher dimensional space, all contained in a bounded domain, and no two of the sets have a point in common, the measure of $G_1 + G_2 + \dots$ is the limiting sum of the measures of the sets.

That $G_1 + G_2 + \dots$ is measurable is a particular case of the last theorem.

In accordance with what has been shewn above $m\,(G_1 + G_2 + \dots)$ differs from the measure of the set consisting of all the sets $\alpha_1, \alpha_2', \alpha_3', \dots$ by less than ϵ. Also $G_2, G_3, \dots$ are all parts of $C\,(G_1)$; so that G_1 is enclosed in α_1, G_2 in α_2', G_3 in α_3', ...; hence $m\,(G_1)$ differs from $m\,(\alpha_1)$ by less than $\frac{1}{2}\epsilon$, $m\,(G_2)$ differs from $m\,(\alpha_2')$ by less than $\frac{1}{2^2}\,\epsilon$, and in general $m\,(G_i)$ differs from $m\,(\alpha_i')$ by less than $\epsilon/2^i$. Therefore $\sum_{i=1}^{\infty} m\,(G_i)$ differs from

$$m\,(\alpha_1) + m\,(\alpha_2') + \dots + m\,(\alpha_i') + \dots$$

by less than ϵ. It follows that $\sum_{i=1}^{\infty} m\,(G_i)$ and $m\,(G_1 + G_2 + \dots)$ differ from one another by less than 2ϵ, and, since ϵ is arbitrarily small, the theorem is established.

It should be observed that although, for each set G_i, a pair of sets of intervals α_i, β_i can be so determined that the measure of $D\,(\alpha_i, \beta_i)$ is less than the prescribed number $\epsilon/2^i$, there are an infinite number of such pairs α_i, β_i which satisfy this condition, and one such pair is selected out of this infinite number. It is however assumed, in the above proof, that an indefinite number of such selections is made, namely one for each value of i. It may not be possible to give a law which determines how this selection is to be made for every value of i, unless the given sets are of some particular nature such that we are able to assign such a law. Consequently, the existence has been assumed of the sets α_i, β_i, for $i = 1, 2, 3, \dots$, independently of whether we are able to define these sets or not. This assumption is a particular case of an axiom which will be discussed in Chapter IV, and which is known as Zermelo's axiom, or as the general principle of selection. A corresponding remark is applicable to another proof of the above theorem which has been given* by de la Vallée Poussin. This proof depends upon the fact that a measurable set G_n is the sum of a closed set F_n and of a set e_n, of which the exterior measure is less than an arbitrarily small number ϵ_n. For an enumerable sequence of sets, the choice of the sets F_n requires the use of Zermelo's axiom (see § 197).

131. *If a measurable set G_1 contains another measurable set G_2, the set $G_1 - G_2$ is measurable, and its measure is $m\,(G_1) - m\,(G_2)$.*

The complement of $G_1 - G_2$ consists of G_2 together with $C\,(G_1)$, hence $G_1 - G_2$ is measurable. Further, since $G_1 = (G_1 - G_2) + G_2$, we have

$$m\,(G_1) = m\,(G_1 - G_2) + m\,(G_2).$$

---

* See *Intégrales de Lebesgue*, pp. 22–25.

If $G_1, G_2, \dots G_n, \dots$ are all measurable sets, the set $D(G_1, G_2, \dots G_n, \dots)$ of points common to all the given sets is also measurable.

For the complement of $D(G_1, G_2, \dots)$ is $M\{C(G_1), C(G_2), \dots\}$; and since $C(G_1), C(G_2), \dots$ are all measurable, it follows that the complement of $D(G_1, G_2, \dots)$ is measurable, and therefore that the set itself is measurable.

If H is the set of points each of which belongs to an infinite number of the measurable sets $G_1, G_2, G_3, \dots$, the set H is measurable*

For the set $C(H)$ consists of those points which belong to none, or only to a finite number, of the sets $G_1, G_2, \dots$; and hence $C(H)$ consists of the points belonging to one or more of the sets $L_1, L_2, \dots L_n, \dots$, where L_n denotes the set $D\{C(G_n), C(G_{n+1}), \dots\}$. The sets L_n are all measurable, and hence $C(H)$ is measurable; and therefore H is measurable.

If K is the set of points each of which belongs to all the measurable sets $G_n, G_{n+1}, \dots$, where n has a definite value for each point of K, the set K is measurable.*

For the set $C(K)$ is the set of points each of which belongs to an infinite number of the measurable sets $C(G_1), C(G_2), \dots$, and hence, by the last theorem, $C(K)$ is measurable. Therefore K is a measurable set.

If G_ω is the inner limiting set of a sequence $\{G_n\}$ of measurable sets G_n, each of which contains the next, $m(G_\omega) = \lim\limits_{n\sim\infty} m(G_n)$.

For $$(G_1 - G_2) + (G_2 - G_3) + \dots = G_1 - G_\omega$$

and $$m(G_n - G_{n+1}) = m(G_n) - m(G_{n+1}),$$

hence $$m(G_1) - \lim_{n\sim\infty} m(G_n) = m(G_1 - G_\omega) = m(G_1) - m(G_\omega),$$

and therefore $$m(G_\omega) = \lim_{n\sim\infty} m(G_n).$$

If G_ω is the outer limiting set of a sequence $\{G_n\}$ of measurable sets G_n, each of which is contained in the next, and if G_ω is a bounded set,

$$m(G_\omega) = \lim_{n\sim\infty} m(G_n).$$

For $$G_1 + (G_2 - G_1) + (G_3 - G_2) + \dots = G_\omega$$

and therefore $$m(G_\omega) = \lim_{n\sim\infty} m(G_n).$$

If† $G_1, G_2, \dots G_n, \dots$ be a sequence of sets such that

$$\lim_{n\sim\infty} M(G_n, G_{n+1}, \dots) = \lim_{n\sim\infty} D(G_n, G_{n+1}, \dots),$$

then, if either of these limits be denoted by G_ω, $m(G_\omega) = \lim\limits_{n\sim\infty} m(G_n)$.

---

* See Borel's *Leçons sur les fonctions de variables réelles*, p. 18. The set H is named by Borel the "ensemble limite complet," and the set K the "ensemble limite restreint" of the given sequence of sets.

† See de la Vallée Poussin, *Trans. Amer. Math. Soc.* vol. XVI (1915), p. 437.

The sets $M(G_n, G_{n+1}, \ldots)$ are such that each contains the next, and therefore $m(G_\omega) = \lim_{n\sim\infty} m[M(G_n, G_{n+1}, \ldots)] \geq \lim_{n\sim\infty} m(G_n)$; and the sets

$$D(G_n, G_{n+1}, \ldots)$$

are such that each is contained in the next, and therefore

$$m(G_\omega) = \lim_{n\sim\infty} m[D(G_n, G_{n+1}, \ldots)] \leq \lim_{n\sim\infty} m(G_n);$$

we thus conclude that $m(G_\omega) = \lim_{n\sim\infty} m(G_n)$.

In general $\lim_{n\sim\infty} M(G_n, G_{n+1}, \ldots)$ is called the upper limit of G_n, and may be denoted by $\overline{\lim_{n\sim\infty}} G_n$; and $\lim_{n\sim\infty} D(G_n, G_{n+1}, \ldots)$ is called the lower limit of G_n, and may be denoted by $\underline{\lim_{n\sim\infty}} G_n$. In the case contemplated in the theorem, G_n has a unique limit denoted by $\lim_{n\sim\infty} G_n$, or G_ω.

SETS THAT ARE MEASURABLE (B)

132. All the sets which have been shewn to be measurable, in accordance with the definition of a measurable set given in § 128, are obtained from the single point, the single interval, or cell, open or closed, by taking a finite, or enumerably infinite, set of such fundamental sets, and by taking the set common to a finite, or an enumerably infinite, number of the sets so obtained, or by taking the complements of the sets so obtained. All sets defined in this manner are said to be measurable (B), since they were the only kind of measurable sets contemplated by Borel in his original treatment of metric properties. Thus the complement of any set that is measurable (B) is so also; the set which is common to any finite, or enumerably infinite, number of sets all measurable (B) is also measurable (B); also closed and completely open sets are all measurable (B). It can be shewn however that measurable sets may exist which are not measurable (B). For example, any component whatever of a perfect set, of measure zero, has its external measure zero, and is therefore measurable. The question whether a valid definition can be given of a set which is not measurable will not be here discussed, as the question of the validity of such definitions depends upon debatable questions in the abstract theory of aggregates. The relation of measurable sets in general to those which are measurable (B) is contained in the following theorem:

Any measurable set G is contained in a set G_1 which is measurable (B), *and such that $m(G_1) = m(G)$; and G contains a set G_2 which is measurable* (B), *and such that $m(G_2) = m(G)$.*

If $\{\epsilon_n\}$ denote a sequence of decreasing positive numbers converging to zero, the set G can be enclosed in a set of non-overlapping intervals, or open domains, $\{\alpha_n\}$ of total measure $< m(G) + \epsilon_n$.

The inner limiting set G_1 of the sequence of sets of intervals, or domains, is measurable (B), and its measure is $m(G)$, as has been shewn in § 131; also G_1 contains G as a component. The set $G_1 - G$ has measure zero, and its points can be enclosed in a set of intervals, or domains, β_n, contained in α_n, and therefore of measure $< \epsilon_n$. The inner limiting set H, of all the sets β_n, is measurable (B), and its measure is zero; and the set $G - H \equiv G_2$ is measurable (B), and has $m(G)$ for its measure. Moreover G_2 is contained in G.

A classification of sets of measure zero has, on account of the fundamental importance of such sets in the theory of functions, been undertaken* by Borel.

CONGRUENT SETS

133. The definition of the measure of a set, given in § 128, having been shewn to apply to the class of measurable sets, it has been shewn that, so far as such measurable sets are concerned, we have a metric system which satisfies the postulates (1), (2), (3), (4), of § 120, namely that the measure of such a set is a completely additive system, and that it is in agreement with the elementary theory of the measure of lengths, areas, and volumes, of intervals and cells. It only remains to be shewn that (5), of § 120, is satisfied by the definition which has been introduced and developed.

A congruent transformation of a set of points is, expressed in geometrical language, equivalent to a translation together with a rotation. That a translation does not affect the metric properties is obvious from the fact that it only consists of an addition of fixed numbers $h_1, h_2, \dots h_p$ to the numbers $x_1, x_2, \dots x_p$ which represent a point; and in all the definitions and proofs connected with the metric properties of a set, only the differences $x_1' - x_1, x_2' - x_2, \dots$ for pairs of points are involved. So far as a transformation by rotation is concerned, it is only necessary to shew that a cell when rotated has a measure which is equal to that of the original cell. It will be observed that all the cells employed in the descriptive properties of sets, and so far in the metric theory, are orientated alike. We cannot therefore, *ab initio*, consider the set of points of a cell, as having its measure unaltered when the cell is rotated, and when it is therefore no longer a cell in the original sense of the term. For simplicity the case of the rectangle, or plane cell, will here be considered.

The original cell may be taken to be the set of points (x, y) for which

$$-a \leqq x \leqq a, \quad -b \leqq y \leqq b.$$

Corresponding to (x, y) we have in the congruent set, obtained by rotation through an angle α, $x' = x\cos\alpha + y\sin\alpha$, $y' = -x\sin\alpha + y\cos\alpha$.

---

[*] See *Bull. de la soc. math. de France*, vol. XLI (1913), p. 1.

Thus the points of the cell correspond to the points of the three sets defined by

$$-\sec\alpha\,(a+y\sin\alpha)\leqq x\leqq\operatorname{cosec}\alpha\,(b+y\cos\alpha)$$
$$-a\sin\alpha-b\cos\alpha\leqq y\leqq -a\sin\alpha+b\cos\alpha,$$
$$\operatorname{cosec}\alpha\,(y\cos\alpha-b)\leqq x\leqq\operatorname{cosec}\alpha\,(y\cos\alpha+b)$$
$$-a\sin\alpha+b\cos\alpha\leqq y\leqq a\sin\alpha-b\cos\alpha,$$
$$b\cos\alpha-a\sin\alpha\leqq y\leqq a\sin\alpha+b\cos\alpha$$
$$-\operatorname{cosec}\alpha\,(b-y\cos\alpha)\leqq x\leqq\sec\alpha\,(a-y\sin\alpha),$$

provided $\tan\alpha > b/a$.

In the first of the above sets, divide the interval of y into n equal parts; and consider the part for which y is in one of these parts $(-c_r, -c_{r+1})$, where $c_r < c_{r+1}$. The corresponding part of the set includes the rectangle of which the measure is

$$(c_{r+1}-c_r)\,\{a\sec\alpha+b\operatorname{cosec}\alpha+c_r(\tan\alpha+\cot\alpha)\},$$

and is included in the rectangle of which the area is

$$(c_{r+1}-c_r)\,\{a\sec\alpha+b\operatorname{cosec}\alpha+c_{r+1}(\tan\alpha+\cot\alpha)\}.$$

The difference of these is $(c_{r+1}-c_r)^2(\cot\alpha+\tan\alpha)$.

The sum of all such differences is less than $n^{-2}(\cot\alpha+\tan\alpha)\,4b^2\cos^2\alpha$, which can be made arbitrarily small by sufficiently increasing n. The measure of the set is then easily shewn to be $2b^2\cot\alpha$. Similarly it can be shewn that the measures of the other two sets are

$$(a\sin\alpha-b\cos\alpha)\,4b\operatorname{cosec}\alpha,\ \text{ and }\ 2b^2\cot\alpha;$$

thus the sum of the measures of the three parts is $4ab$, which is the measure of the original cell.

The case in which $\tan\alpha$ has a value $< b/a$ may be considered in a similar manner. The method here adopted is clearly capable of extension to the case of a cell in any number of dimensions.

THE MEASURE OF UNBOUNDED SETS

134. Let G be an unbounded set of p dimensions, and let it be such that the component of G which is in any finite p-dimensional cell is measurable. Consider a sequence of such cells, each of which is contained in the next, and such that if $(a_n^{(1)}, a_n^{(2)}, \dots a_n^{(p)};\ b_n^{(1)}, b_n^{(2)}, \dots b_n^{(p)})$ is the nth cell of the sequence, $a_n^{(1)}, a_n^{(2)}, \dots a_n^{(p)}$ all have the limit $-\infty$, and $b_n^{(1)}, b_n^{(2)}, \dots b_n^{(p)}$ all have the limit $+\infty$, as n is indefinitely increased. If G_n be the component of G in this cell, the sequence of numbers $m(G_n)$ is non-diminishing. Consequently $m(G_n)$ either converges to a finite limit, or becomes indefinitely great, as n is indefinitely increased. If $\lim\limits_{n\sim\infty} m(G_n)$ is finite, the set G is regarded as measurable, and $m(G)$ is defined to be equal to the limit. If the limit is indefinitely great, G is measurable, but of infinite measure. If H be any bounded measurable set of points, it is contained in the nth cell, if n be sufficiently great. The set $D(G, H)$ is the same as $D(G_n, H)$, and this set is measurable. Also $m[D(G, H)] \leq m(G_n)$. If H be identified successively with the sets H_r of a sequence, such that H_r is contained in H_{r+1}, we have

$$m[D(G, H_r)] \leq m(G_{n_r}) \leq m(G),$$

where n_r is an integer corresponding to r. The sequence of numbers $m[D(G, H_r)]$ is monotone, non-decreasing, and has a limit $\leq m(G)$. This holds good, even if $\lim_{r\sim\infty} m(H_r)$ is indefinitely great. Thus the part of G that is contained in any measurable set H is measurable, if G be so, and its measure does not exceed $m(G)$.

The theorem $m(G_1) + m(G_2) + \ldots = m(G_1 + G_2 + \ldots)$, where $G_1, G_2, \ldots$ are measurable sets, no two of which have a point in common, and such that $G_1 + G_2 + \ldots$ has a finite measure, holds good when $G_1, G_2, \ldots$ are unbounded measurable sets, of finite measure. For, if Δ_n denote the cell employed above, we have

$$m[D(\Delta_n, G_1)] + m[D(\Delta_n, G_2)] + \ldots = m[D(\Delta_n, G_1 + G_2 + \ldots)].$$

If r be so large that $m(G_{r+1}) + m(G_{r+2}) + \ldots < \epsilon$, we have also

$$m[D(\Delta_n, G_{r+1})] + m[D(\Delta_n, G_{r+2})] + \ldots < \epsilon,$$

for every value of n.

Hence, we have

$$m[D(\Delta_n, G_1)] + m[D(\Delta_n, G_2)] + \ldots + m[D(\Delta_n, G_r)] + \theta_n\epsilon$$
$$= m[D(\Delta_n, G_1 + G_2 + \ldots)],$$

where θ_n is such that $0 < \theta_n < 1$.

It follows, by letting n increase indefinitely, that

$$m(G_1) + m(G_2) + \ldots + m(G_r) + \bar{\theta}\epsilon = m(G_1 + G_2 + \ldots),$$

where $\bar{\theta}$ is such that $0 \leq \bar{\theta} \leq 1$. Since ϵ is arbitrary, it follows that $m(G_1) + m(G_2) + \ldots$ converges to $m(G_1 + G_2 + \ldots)$.

It is easily seen that, *if the unbounded measurable set G is the outer limiting set of a sequence $\{G_n\}$ of bounded measurable sets, each of which is contained in the next, then $m(G) = \lim_{n\sim\infty} m(G_n)$.*

THE MEASURE OF SETS RELATED TO A SYSTEM OF SETS

135. It has been shewn in § 131 that, for a sequence $\{G_n\}$ of measurable sets, each of which contains the next, if $m(G_n)$ is, for every value of n, greater than a fixed positive number C, the measure of the inner limiting set is $\geq C$.

In case the sets $G_1, G_2, \ldots$, each of which contains the next, are not assumed to be measurable, the corresponding result holds as regards the interior measures of the sets. We can obtain the following theorem:

If[*] $G_1, G_2, \ldots$ *be a sequence of sets of points, each of which contains the next, and if the interior measure of each set is greater than a fixed positive number* C, *then the interior measure of the inner limiting set is* $\geq C$.

Closed components, $P_1, Q_2, Q_3, \ldots$ of the sets $G_1, G_2, G_3, \ldots$ can be so determined that

$$m(P_1) > m_\iota(G_1) - \frac{1}{2}\epsilon, \quad m(Q_2) > m_\iota(G_2) - \frac{1}{2^2}\epsilon, \quad m(Q_3) > m_\iota(G_3) - \frac{1}{2^3}\epsilon, \ldots;$$

where ϵ is an arbitrarily chosen positive number. The set Q_2 has a closed component $P_2 \equiv D(P_1, Q_2)$, of measure

$$m(P_2) = m(P_1) + m(Q_2) - m[M(P_1, Q_2)].$$

Since $M(P_1, Q_2)$ is a closed set contained in G_1, its measure does not exceed $m_\iota(G_1)$, hence we have

$$m(P_2) \geq m(P_1) + m(Q_2) - m_\iota(G_1) \geq m_\iota(G_2) - \left(\frac{1}{2} + \frac{1}{2^2}\right)\epsilon.$$

Next, by considering $P_3 \equiv D(P_2, Q_3)$, it can be shewn as before that

$$m(P_3) \geq m_\iota(G_3) - \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3}\right)\epsilon;$$

and so on. We have now a sequence of closed sets $P_1, P_2, P_3, \ldots$. each of which contains the next, and such that the measure of each of them is $> C - \epsilon$. Therefore the measure of P_ω, their inner limiting set, is $\geq C - \epsilon$; and P_ω is a component of G_ω. It follows, since ϵ is arbitrarily small, that $m_\iota(G_\omega) \geq C$.

136. We are now in a position to establish the following general theorem:

If[*] $G_1, G_2, \ldots G_n, \ldots$ *be a sequence of sets of points, each of which sets is a component of a closed set, of finite content* l, *and if the interior measure of each of the sets* $G_1, G_2, \ldots G_n, \ldots$ *is greater than a fixed number* C, *then there exists a set of points, of interior measure* $\geq C$, *and of the power of the continuum, such that each point of the set belongs to an infinite number of the given sets.*

Choose a closed component of each of the given sets, of content $> C$; let these components be $Q_1, Q_2, \ldots Q_n, \ldots$. Choose an integer m such that $mC \leq l < (m + 1)C$, and let us consider the first n ($> m + 1$) of the sets $Q_1, Q_2, \ldots$. The points common to any pair of these closed sets form a closed set, and the set which contains all the points which belong to at least two of the n closed sets is also a closed set $Q_{1,n}$ of content I_{1n}. Those points of $Q_{1,n}$ which belong to Q_1 form a closed set of content $\leq I_{1n}$; hence there is a set of points of Q_1, of measure $\geq I(Q_1) - I_{1n}$, which do not belong to any of the sets $Q_2, Q_3, \ldots Q_n$; and the measure of this set is $> C - I_{1n}$. Similarly, each of the sets $Q_2, Q_3, \ldots Q_n$ has a component of measure $> C - I_{1n}$, consisting of points which do not belong to any of the other sets, or to Q_1. The measure of all these sets added together is $> n(C - I_{1n})$; and must be less than l, since the sets do not contain any points common to two of them, and they are all enclosed in a set of measure l. Hence

---

[*] W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. II (1905), p. 25.

$$n(C - I_{1n}) < (m+1)C, \text{ or } I_{1n} > \left(1 - \frac{m+1}{n}\right)C.$$

It has thus been shewn that the closed set $Q_{1,n}$ has the power of the continuum, since its content is proved to be positive; and this holds for every value of n which is $> m + 1$. Considering now the next n sets $Q_{n+1}, Q_{n+2}, \dots Q_{2n}$, there is a closed set of content $> \left(1 - \frac{m+1}{n}\right)C$, consisting of points each of which belongs to two at least of the sets; and a similar result holds for each system of n sets $Q_{rn+1}, Q_{rn+2}, \dots Q_{(r+1)n}$.

We have now an infinite sequence of closed sets

$$Q_{1,n}, Q_{2,n}, Q_{3,n}, \dots Q_{r,n}, \dots$$

each of which has content $> \left(1 - \frac{m+1}{n}\right)C$, and the points of each of them belong to two at least of the given sets. By applying similar reasoning, and taking n' sets at a time, we see that there are an infinite number of sets each of content $> \left(1 - \frac{m+1}{n}\right)\left(1 - \frac{m+1}{n'}\right)C$, and such that each point of any one of them belongs to four at least of the given sets. Proceeding in this manner, we obtain sets of points, each of content

$$> \left(1 - \frac{m+1}{n}\right)\left(1 - \frac{m+1}{n'}\right)\left(1 - \frac{m+1}{n''}\right)\dots\left(1 - \frac{m+1}{n^{(s)}}\right)C,$$

and such that each point of each set belongs to at least 2^{s+1} of the given sets. Now let $n, n', \dots n^{(s)}$ be so chosen, that

$$\frac{m+1}{n} < \frac{1}{2}\epsilon, \quad \frac{m+1}{n'} < \frac{1}{4}\epsilon, \dots \frac{m+1}{n^{(s)}} < \frac{1}{2^{s+1}}\epsilon;$$

then the content of each of the sets which contains points belonging to 2^{s+1} at least of the given sets is

$$> C\left(1 - \frac{1}{2}\epsilon\right)\left(1 - \frac{1}{4}\epsilon\right)\dots\left(1 - \frac{1}{2^{s+1}}\epsilon\right) > C\left\{1 - \epsilon\left(\frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{s+1}}\right)\right\}$$
$$> C(1 - \epsilon).$$

The process can be carried on without limit; and we see that the set which consists of all points belonging to 2^{s+1} at least of the given sets contains closed components of content $> C(1 - \epsilon)$. Considering the sequence $P_1, P_2, \dots$ of sets such that P_1 contains all points that belong to two at least of the given sets, P_2 contains all points that belong to 2^2 at least of the given sets, and so on, it is clear that P_1 contains P_2, and

P_2 contains P_3, etc. But the interior measure of each set is $> C(1-\epsilon)$; hence, in accordance with the theorem of § 135, there exists a set of points common to all the sets $P_1, P_2, \ldots$, of interior measure $\geqq C(1-\epsilon)$. This set consists of points which belong each to an infinite number of the given sets; and its interior measure is $\geqq C$, since ϵ is arbitrarily small. The set has the power of the continuum, since it contains closed components of content greater than zero. It should be observed that the employment of the sets $\{Q_n\}$ involves an infinite number of acts of choice.

The theorem that has been now established is of considerable importance on account of the applications of it which can be made in various parts of the theory of functions, it is due* to W. H. Young. That particular case of the theorem in which the sets are all measurable was first stated†, without proof, by Borel.

An important case of the theorem arises if we suppose each of the sets to consist of a finite, or an enumerably infinite, set of closed intervals; in which case the sets are all measurable. The theorem may then be stated as follows:

If there be given an infinite number of sets of intervals, in a finite segment, each set consisting of a finite, or an enumerably infinite, number of non-overlapping intervals, and if the measure of each set of intervals is greater than some fixed positive number C, then there exists a set of points, having the power of the continuum, and of measure $\geqq C$, such that each point of the set belongs to an infinite number of the given sets of intervals.

This theorem contains the completion, and generalization, of a theorem due to Arzelà‡ which is stated by him as follows:

Let y_0 be a limiting point of any set of numbers (y), and let

$$G_0 = (y_1, y_2, \ldots)$$

be a sequence of numbers of (y) which converges to the limit y_0. Assuming the variables to be orthogonal coordinates of a point in a plane, let the set of straight lines $y = y_1$, $y = y_2$, $y = y_3, \ldots$, be drawn, and let a set of intervals be taken on the portion of each of these straight lines which is in the interval (a, b), of x. Suppose that each set of intervals is finite in number, and that this number is variable from one straight line to another, but increases indefinitely with the index s, of y_s. Let the sum of the intervals $\delta_{1,s}, \delta_{2,s}, \ldots \delta_{n,s}$ on the line $y = y_s$, be d_s. If for every value of s, d_s is greater than C, a determinate positive number, there necessarily exists at least one point x_0 in the interval (a, b), such that the straight line $x = x_0$ intersects an infinite number of the intervals δ.

---

* *Proc. Lond. Math. Soc.* (2), vol. II (1904), p. 26.

† *Comptes Rendus*, vol. CXXXVII (1903), p. 966.

‡ *Rend. dell' Ac. dei Lincei*, (4), vol. I (1885), p. 637; a second proof, which is however not rigorous, has been given by Arzelà in the *Memorie della R. Acc. d Sc. di Bologna*, (5), vol. VIII (1899).

Arzelà subsequently removed the condition that each set of intervals is a finite one.

VITALI'S THEOREM

136^1. An important theorem has been given* by Vitali which may be regarded as an extension, in relation to metric properties, of the descriptive theorem of Heine-Borel and its extensions, discussed in §§ 73–77.

The following preliminary theorem will be first established:

Let C denote a set of cells in p dimensions, not necessarily enumerable, such that each cell has its sides of equal length, and such that the measures of the cells have a finite upper boundary k_1. Let it be assumed that each cell of C has at least one point in common with each cell of some non-finite set of cells, all belonging to C. Let H be the set of points equivalent to C; and let it be assumed that H has its measure finite. Then a finite set of cells exists, all belonging to C, such that no two cells of the finite set have a point in common, and such that the sum of their measures is greater than

$$\frac{1}{3^p} m(H) - \epsilon,$$

where ϵ is a prescribed positive number.

A cell c_1, belonging to C, exists such that $m(c_1) > k_1 - \frac{1}{2^2}\epsilon$; let C_1 denote the part of C such that each cell of C_1 has at least one point in common with c_1; such part exists by hypothesis. The points of the set L_1, equivalent to C_1, that do not belong to c_1 have a measure $\leq (3^p - 1) k_1$. Let k_2 be the upper boundary of the measures of the cells $C - C_1$; then there exists in $C - C_1$ a cell c_2, which necessarily has no point in common with c_1, and of which the measure is $> k_2 - \frac{1}{2^3}\epsilon$. Let C_2 denote the part of $C - C_1$ of which each cell has at least one point in common with c_2. The set of points L_2 equivalent to C_2 that do not belong to c_2 has measure $\leq (3^p - 1) k_2$. Proceeding in this manner, we form a sequence $c_1, c_2, c_3, \ldots$ of cells such that their total measure is not greater than $m(H)$; and therefore

$$\sum_{n=1}^{\infty} \left(k_n - \frac{\epsilon}{2^{n+1}}\right) < m(H).$$

Since $\sum_{n=1}^{\infty} k_n$ is accordingly convergent, we have $\lim_{n\sim\infty} k_n = 0$; it then follows that $C = \sum_{n=1}^{\infty} C_n$, for if a cell c existed in C which for no value of n belongs to C_n, every value of k_n would be $\geq m(c)$, and thus k_n would not converge to zero. Let $\bar{C}$ denote the sum of $c_1, c_2, \ldots c_n, \ldots$, and let $\bar{L}$ denote the set of points equivalent to the totality of the sets $L_1, L_2, \dots$. Then $H = \bar{C} + \bar{L}$; also $m(\bar{C}) < \sum_{n=1}^{\infty} k_n$, $m(\bar{L}) \leqq (3^p - 1) \sum_{n=1}^{\infty} k_n$, and therefore

---

* *Atti d. R. Accad. d. Torino*, vol. XLIII (1908), p. 229. A somewhat different, and very general, statement of the theorem is given by Carathéodory, *Vorlesungen über reelle Funktionen* (1918), p. 299.

$$m(H) < 3^p \sum_{n=1}^{\infty} k_n.$$

Also we have $m(\bar{C}) > \sum_{n=1}^{\infty} k_n - \tfrac{1}{2}\epsilon$, and therefore

$$m(\bar{C}) > \frac{1}{3^p} m(H) - \tfrac{1}{2}\epsilon.$$

An integer n_1 can be so chosen that $m(c_1 + c_2 + \dots + c_{n_1}) > m(\bar{C}) - \tfrac{1}{2}\epsilon$; then $m(c_1) + m(c_2) + \dots + m(c_{n_1}) > \frac{1}{3^p} m(H) - \epsilon$: thus the required cells $c_1, c_2, \dots c_{n_1}$ exist.

The main theorem given by Vitali is the following:

Let C denote a set of cells in p dimensions, each of which has sides of equal length. Let $C^{(\eta)}$ denote that part of C which consists of cells each of which has its measure $< \eta$, such part being assumed to exist for each positive value of η; and let G be the inner limiting set of the sets of points equivalent to $C^{(\eta_1)}, C^{(\eta_2)}, \dots C^{(\eta_n)}, \dots$, where $\{\eta_n\}$ is a diminishing sequence of numbers converging to zero; and let G be assumed to have finite measure. Then an enumerable, or a finite, set of cells exists, all belonging to C, such that no two of them have a point in common, and such that their total measure is $\geqq m(G)$.

Since G has finite measure, the set of points equivalent to $C^{(\eta)}$ must have finite measure, provided η is sufficiently small. There exists therefore a part C_1, of C, consisting of a part $C^{(\eta)}$, for some value of η, such that, if H_1 is the set of points equivalent to C_1, $m(H_1)$ has finite measure. In accordance with the last theorem a set $c_1, c_2, \dots c_{n_1}$, of cells belonging to C_1 exists such that no two of them have a point in common, and such that

$$\sum_{r=1}^{r=n_1} m(c_r) > \frac{1}{3^p} m(H_1) - \tfrac{1}{2}\epsilon > \frac{1}{3^p} m(G) - \tfrac{1}{2}\epsilon.$$

First, let it be assumed that every cell of C_1 has a point in common with one at least of the cells $c_1, c_2, \dots c_{n_1}$. If, for some sufficiently small value of ζ, $C^{(\zeta)}$ contains no cells partly outside the cells $c_1, c_2, \dots c_{n_1}$, all the points of G must be in one of the cells of this finite set; and in that case

$$\sum_{r=1}^{r=n_1} m(c_r) > m(G),$$

and the set required has been obtained. If, for every value of ζ, however small, there are cells of $C^{(\zeta)}$ partly outside the cells c_r, all these cells must be contained in neighbourhoods of the cells c_r whose measures diminish indefinitely with ζ; therefore there can exist no points of G exterior to all the cells c_r. Hence, as before $\sum_{r=1}^{r=n_1} m(c_r) > m(G)$, and $c_1, c_2, \ldots c_{n_1}$ is the required set.

Next, let it be assumed that a set C_2, a part of C_1, exists such that no cell of C_2 has a point in common with any of the cells $c_1, c_2, \ldots, c_{n_1}$. Let the set G_2 be the inner limiting set of the sequence $\{C_2^{(\eta_n)}\}$, where $C_2^{(\eta_n)}$ consists of those cells of C_2 whose measure is $< \eta_n$; and let H_2 be the set of points equivalent to C_2. We have then the relation

$$m(G_2) \geqq m(G) - \sum_{r=1}^{r=n_1} m(c_r).$$

In C_2 a finite set of intervals $c_{n_1+1}, c_{n_1+2}, \ldots, c_{n_2}$ exists, no two of which have a point in common, and such that

$$\sum_{r=n_1+1}^{r=n_2} m(c_r) > \frac{1}{3^p} m(H_2) - \frac{1}{2^2}\epsilon > \frac{1}{3^p} m(G_2) - \frac{1}{2^2}\epsilon.$$

In case every cell of C_2 has a point in common with one at least of the cells $c_{n_1+1}, c_{n_1+2}, \ldots, c_{n_2}$, it can be shewn as before that

$$\sum_{r=n_1+1}^{r=n_2} m(c_r) > m(G_2);$$

then we have from the relation given above

$$\sum_{r=1}^{r=n_2} m(c_r) > m(G),$$

and thus the required set of cells exists. If however, there exists a set C_3 which is a part of C_2 such that no cell of C_3 has a point in common with any of the cells $c_{n_1+1}, c_{n_1+2}, \ldots c_{n_2}$, and G_3 be defined as in the preceding definition of G_2, we have

$$m(G_3) \geqq m(G_2) - \sum_{r=n_1+1}^{r=n_2} m(c_r).$$

We can proceed indefinitely in this manner, or until the process ceases at some stage. We obtain then, either a finite set of cells having the required property, or else a sequence of cells $\{c_r\}$. The series $\sum_{r=1}^{\infty} m(c_r)$ is convergent, since $\sum_{r=1}^{r=n_p} m(c_r) < m(H_1)$; consequently the series

$$\sum_{r=1}^{\infty} \left(\frac{1}{3^p} m(G_r) - \frac{\epsilon}{2^r}\right)$$

is convergent, where $G_1 \equiv G$; and therefore, the series $\sum_{r=1}^{\infty} m(G_r)$ being convergent, we have $\lim_{r\sim\infty} m(G_r) = 0$. Since

$$m(G_r) \geqq m(G) - \sum_{s=1}^{s=n_r} m(c_s),$$

we now have, as $r \sim \infty$, $\sum_{s=1}^{\infty} m(c_s) \geqq m(G)$. The set $\{c_s\}$, of cells, accordingly satisfies the conditions of the theorem.

It is clear that, in the general case, the solution of the set $\{c_n\}$, in the foregoing theorem, requires an application of the multiplicative axiom, but in particular cases, in which the definition of the set C, of cells, is of a special character, the use of the axiom may be dispensed with, by the assignment of a law of selection.

Important applications of Vitali's theorem can be made in the case when, starting with a given measurable set of points K, with each point P, of K, a set of cells, with P as the centre of each cell (with equal sides), is defined, and so that the lower limit of the measures of the cells is zero; the set may consist either of an enumerable sequence $\Delta_n(P)$, such that

$$\lim_{n\sim\infty} m[\Delta_n(P)] = 0,$$

or of a continuous set consisting of all such cells of which the measure has a finite upper limit. The set C, in the theorem, will then be taken to be the totality of the sets of cells corresponding to all the points P, of the given set K. The inner limiting set G will then consist of the set K, possibly increased by the addition of a set, all the points of which are limiting points of K. This has been shewn to be the case for linear sets in § 98, and the reasoning there given can be at once adapted to the case of sets in any number of dimensions. In case K is a closed set, we have $G \equiv K$. If K is an open set, and all the cells C are interior to K, the set G can contain no points that do not belong to K; and since Σc_r cannot have its measure greater than that of K, it follows from the theorem that the sum of the measures of all the cells c_r is equal to $m(K)$. Hence the set of cells $\{c_r\}$ contains every point of K, with the exception of a set of which the measure is zero. We have accordingly the following theorem:

If, at each point P of an open set K, of finite measure, a set of cells, each with equal sides, is defined, all the cells being interior to K, and the lower limit of their measures being zero, an enumerable set of cells exists, all belonging to the complete set so defined, and no two of which have a point in common, which contains all the points of the open set K, except those belonging to a set of measure zero. The set with centre P may form a sequence, or be unenumerable.

The following corollary from Vitali's theorem is useful in applications:

If E be a measurable set of points, of finite measure, a finite set of cells can be determined, no two of which have a point in common, which includes all the points of a part of E of measure $> m(E) - \eta$, where η is an assigned positive number, and such that the sum of the measures of the cells is $< m(E)$.

The set E contains a closed set H, of measure $> m(E) - \frac{1}{2}\eta$ and $< m(E)$. Taking the set C of cells to be interior to an open set K of measure $< m(E)$ containing H (see § 127), C is equivalent to a set of points that has measure $< m(E)$, and the inner limiting set G consists, as above, of almost all the points of K. The enumerable set $\{c_r\}$ contains almost all the points of K, and therefore almost all the points of H; and its total measure is $m(K) < m(E)$. Hence a finite set $c_1, c_2, \dots c_s$ of these cells exists the sum of whose measures is $< m(E)$ and it contains a part of E of measure $> m(E) - \eta$.

THE METRIC DENSITY OF A SET OF POINTS

137. Let P be any point in linear, plane, or p-dimensional space, and let G be a measurable set of points in that space. The point P may, or may not, belong to G. Suppose the interval, or cell, Δ_h to have the point P as its centre, and to have all its sides of equal length h; let G_{Δ_h} be the part of G contained in Δ_h.

If a definite number $\rho_P(G)$ exists such that

$$\left| \rho_P(G) - \frac{m(G_{\Delta_h})}{m(\Delta_h)} \right| < \epsilon,$$

where ϵ is an arbitrarily chosen positive number, for all such cells Δ_h for which $h < h_\epsilon$, a number dependent on ϵ, then the number $\rho_P(G)$ is said to be the *metric density* of the set G at the point P. It is clear that, unless P is a limiting point of G, whether it belongs to G, or not, the metric density at P is zero. The definition is equivalent to

$$\rho_P(G) = \lim_{h \sim 0} \frac{m(G_{\Delta_h})}{m(\Delta_h)},$$

whenever the limit exists as a unique number; the cell Δ_h having equal sides, and the point P being at its centre.

It may however happen that such unique limit does not exist, there may be different sequences of values of h, all converging to zero, for which the above ratio converges to different values. There is then an upper, and a lower, limit of the values to which $m(G_{\Delta_h})/m(\Delta_h)$ may converge. These two numbers are both in the closed interval $(0, 1)$, and the greater of them is called the *upper metric density* of G at P, and may be denoted by $\bar{\rho}_P(G)$; the smaller of the two numbers may be called the *lower metric density* of G at P, and may be denoted by $\underline{\rho}_P(G)$. Thus we have

$$\bar{\rho}_P(G) = \overline{\lim_{h \sim 0}} \frac{m(G_{\Delta_h})}{m(\Delta_h)}, \quad \underline{\rho}_P(G) = \underline{\lim_{h \sim 0}} \frac{m(G_{\Delta_h})}{m(\Delta_h)}.$$

At every point P, of G, or G', we have $\bar{\rho}_P(G) \geqq \underline{\rho}_P(G)$; when the sign of equality holds, there is a metric density $\rho_P(G)$ at P.

If $C(G)$ be the complement of G relatively to a cell, or interval, in which G is contained, we see that

$$\frac{m(G_{\Delta_h})}{m(\Delta_h)} + \frac{m\{C(G)\}_{\Delta_h}}{m(\Delta_h)} = 1,$$

and thence it follows that

$$\bar{\rho}_P(G) + \underline{\rho}_P\{C(G)\} = 1, \quad \underline{\rho}_P(G) + \bar{\rho}_P\{C(G)\} = 1;$$

and in case the metric density of G, at P, exists, that of $C(G)$, at P, also exists, and the sum of the two is unity. It is clear that, if $\bar{\rho}_P(G) = 0$, then also $\rho_P(G)$ exists, and has the value zero.

More general definitions of the upper and lower metric densities at a point may be obtained by employing, instead of the special set of cells Δ_h, other more general sets of points which satisfy certain conditions: this was done* by Lebesgue. We shall however make no use of this more general definition.

It is however necessary to employ the conception, due† to de la Vallée Poussin, of the upper and lower metric densities at a point relatively to a system of nets.

If a system of nets $\{D_n\}$ be fitted on to the indefinitely great linear, or p-dimensional, space in which the set G is defined, a point P is defined uniquely by the set $d_1, d_2, \ldots d_n, \ldots$ of meshes of $D_1, D_2, \ldots D_n, \ldots$ which contain it (see § 51). The upper and the lower limits of $\frac{m(G_{d_n})}{m(d_n)}$, as $n \sim \infty$, are defined to be the upper and lower metric densities of G, at P, relative to the system $\{D_n\}$ of nets. When the two have the same value, they define the metric density of G, at P, relative to the system of nets.

138. A set G is said to be *metrically dense* at a point P, whether or not P belongs to G, if, in every cell, or interval, according to the dimensions of the space in which the set is defined, which contains P in its interior, there is a set of points of G of measure > 0. A point at which G is metrically dense must belong to G'.

The metrical density of G may be zero at a point at which G is metrically dense; for $m(G_\Delta)$ may be > 0 for every cell Δ, and yet the limit $m(G_\Delta)/m(\Delta)$ may exist, and have the value zero.

The terms *metrical density*, *metrically dense* are employed because the terms dense and non-dense have already been used in a descriptive sense. A set may be everywhere dense, and yet nowhere metrically dense, as is the case, for example, for the set of rational points in the linear interval $(0, 1)$.

The set of all points at which a set G is metrically dense is closed.

For if, in every neighbourhood of a point P, there are points at which G is metrically dense, it is clear that, at the point P, G is metrically dense.

Those points of a measurable set G at which the set is not metrically dense form a component of G, of measure zero, provided that component exists.

---

* *Annales sc. de l'école normale*, (3), vol. XXVII (1910), p. 387.

† *Intégrales de Lebesgue*, pp. 63, 71; also *Trans. Amer. Math. Soc.* vol. XVI (1915), p. 488.

Apply a system of nets to the cell, or interval, in which G is contained. Let D_{n_1} be the first of the nets which has one or more meshes each of which contains only a part of G of measure zero, let D_{n_1}' denote those meshes of D_{n_1} for which this is the case. Let D_{n_2} ($n_2 > n_1$) be the first net which has meshes, not contained in D_{n_1}', in each of which the component of G has measure zero; let D_{n_2}' denote these meshes. Proceeding in this manner, we define a sequence $D_{n_1}', D_{n_2}', \ldots$ each of which consists of a set of meshes belonging to $D_{n_1}, D_{n_2}, \ldots$ respectively, and such that D_{n_r}' does not belong to $D_{n_{r-1}}'$, for any value of r. Every point P, belonging to G, at which G is not metrically dense belongs to one of the meshes of the sets $D_{n_1}', D_{n_2}', \ldots$, each of which contains a set of points of G of measure zero. Since the set of all points of G contained in all these sets of meshes is of measure zero, it follows that the component of G, at each point of which G is not metrically dense, has measure zero.

A set of points is said to be *metrically dense in itself* when every point of the set is a point at which the set is metrically dense.

If, from any set G, such that $m(G) > 0$, those points at which the set is not metrically dense be removed, there remains a set H such that $m(H) = m(G)$, which is metrically dense in itself. For, in the arbitrarily small neighbourhood of any point of H, there is a set of points of G of measure > 0, and therefore a set of points of H of measure > 0. The following theorem has therefore been established:

Any measurable set, of measure > 0, is the sum of a set of measure zero, which may be non-existent, and of a set that is metrically dense in itself.

In case the set G is closed, we may add to H its limiting points, all of which belong to G; we thus obtain a closed component $\bar{H}$ of G, where $m(\bar{H}) = m(G)$. The set $\bar{H}$ has the same property as H; for, in the neighbourhood of a limiting point of H, there must be a set of points of H of measure > 0. It is clear that $\bar{H}$ has no isolated points, hence it is a perfect set. It has thus been proved that:

A closed set G, of measure > 0, is the sum of a perfect set $\bar{H}$, of measure equal to that of G, and metrically dense in itself, and of a set of measure zero.

Any measurable set G contains a non-dense closed component, of measure arbitrarily less than that of G (§ 125). This non-dense closed component contains a perfect set, also non-dense, which is metrically dense in itself, and of measure equal to that of the component. We have thus proved the following theorem, due to Lusin:

Any measurable set, of measure > 0, contains a non-dense perfect set, metrically dense in itself, and of measure differing by an arbitrarily small amount from that of the given set.

139. With a view to the establishment of the fundamental theorem in the theory of metrical density it is convenient to introduce the conception, due* to de la Vallée Poussin, of conjugate systems of nets.

Let us first consider a net D_1, all the meshes of which are of equal length, d, fixed on to the indefinite interval $(-\infty, \infty)$ on the $x^{(1)}$-axis. Let D_2 be the net obtained by dividing each mesh of D_1 into three equal meshes of breadth $\frac{1}{3}d$; we proceed to obtain D_3, by dividing each mesh of D_2 into three equal parts, and so on indefinitely. We have thus a symmetrical system of nets, such that the breadth of each mesh of D_n is $d/3^{n-1}$. Next consider the net D_1', of which the end-points of the meshes are the middle points of the successive meshes of D_1: from D_1', we form a symmetrical system of nets $\{D_n'\}$ in exactly the same manner as $\{D_n\}$ was formed from D_1. It is then easily seen that the end-points of the successive meshes of D_n' are the middle points of the successive meshes of D_n. The two systems of nets $\{D_n\}, \{D_n'\}$ are then said to be *conjugate* to one another; they have the property that any interval of length $< \frac{1}{2}d$ is contained within a mesh of one at least of the two nets D_n, D_n', provided its length is $< \frac{1}{2}d/3^{n-1}$.

Next, let systems of linear nets, precisely similar to $\{D_n\}$, $\{D_n'\}$, be fitted on to the $x^{(2)}$-axis, and let straight lines of indefinite length, be drawn parallel to the $x^{(1)}$-axis and the $x^{(2)}$-axis, intersecting those axes at the end-points of the meshes of the two pairs of conjugate linear systems of nets. In this manner we obtain four systems of symmetrical nets fitted on to the unbounded plane of $(x^{(1)}, x^{(2)})$; they are said to form four systems of nets, any pair of which are conjugate to one another. Let us consider any square, with its sides parallel to the axis, of side a, $< \frac{1}{2}d$, in the plane; the projections of two sides on the $x^{(1)}$-axis fall within a mesh of at least one of the two nets D_n, D_n', provided $a < \frac{1}{2}d/3^{n-1}$; also the projections of the other sides fall within a mesh of at least one of the two corresponding nets on the $x^{(2)}$-axis. Consequently the square is entirely contained in one at least of the four plane nets of order n, belonging to the four systems of nets which have been fitted on to the plane. The same holds for any cell, provided its greatest side is $< \frac{1}{2}d/3^{n-1}$.

Proceeding in a similar manner, 2^p conjugate systems of nets can be fitted on to the p-dimensional space, and a cell, whose sides are a $(< \frac{1}{2}d)$ is contained within a mesh of at least one of the nets of order n, belonging to the 2^p sets, provided the greatest side of the cell is $< \frac{1}{2}d/3^{n-1}$. We may suppose, as in § 51, that all the meshes of all the nets are semi-closed, so that any point is defined uniquely by the set of meshes of any one of the systems in which it is contained.

---

* *Intégrales de Lebesgue*, p. 64.

The following theorem will be established:

If E and F are two measurable sets of points, in any number of dimensions, and those points of E at which the upper metric density of F is greater than some positive number c, form a component E_1, of E, such that $m(E_1) > 0$, then a system of nets can be so determined that those points of E at which the upper metric density of F relatively to the system of nets is $> c/\lambda$ form a set E_2 such that $m(E_2) > 0$, where λ is a fixed number properly chosen.

Suppose that, at a point P, of E_1, the upper metric density of F is $> c$, then a sequence of cells Δ_h exists, each one of which has its sides equal to h, and with P as its centre, and each cell of the set containing the next, which converges to the point P, and is such that, in each cell, the ratio of the measure of the part of F in the cell to the measure of the cell is $> c$. Consider a set of 2^p conjugate systems of cells, as defined above. For any value of h ($< \frac{1}{2}d$), there exists an integer n, such that

$$\frac{d}{2 . 3^{n-1}} > h \geqq \frac{d}{2 . 3^n}.$$

The cell Δ_h is then interior to a mesh of one of the 2^p conjugate systems, the length of a side of this mesh being $\frac{d}{3^{n-1}}$. The ratio of the measure of the part of F contained in this mesh to the measure of the mesh is $> \frac{c}{6^p}$. Since this is the case for all the values of h in a certain sequence, the upper metric density of F at P, relatively to one at least of the 2^p systems of nets, must be $> \frac{c}{\lambda}$, where λ is a number $> 6^p$. This is the case for each point P belonging to the set E_1, of which the measure is > 0; there must therefore be a part E_2, of E_1, of measure > 0, such that the upper metric density of F, at every point of E_2, relatively to one of the 2^p systems of nets is $> c/\lambda$. That one of the conjugate systems of nets is a system of nets such as is required.

140. The following fundamental theorem will now be established:

The metric density of a measurable set E exists, and is equal to unity, at all points of E, with the possible exception of the points of a component of which the measure is zero; and the metric density of E exists, and is equal to zero, at all points of $C(E)$, with the possible exception of points of a set of measure zero.

We shall first suppose that E is a bounded set. It will be shewn that the points of E at which the upper metric density of $C(E)$ is > 0 form a set of measure zero.

Let us suppose, if possible, that the measure of the component of E, at each point of which the upper metric density of $C(E)$ is greater than a fixed positive number α, is > 0. A system of nets can then be so determined that the upper metric density of $C(E)$, relatively to the system of nets, is $> \alpha/\lambda$, at all points of a component E_α, of E, such that $m(E_\alpha) > 0$. Let Ω be an open set of points which contains E_α, and is such that $m(\Omega) - m(E_\alpha)$ is less than an arbitrarily chosen positive number η.

Any point P, of E_α, is contained in a unique sequence of meshes of the nets, one in each of the nets $D_1, D_2, \ldots D_n, \ldots$. From and after some value s, of n, all these meshes are contained in Ω. Of these meshes $d_s, d_{s+1}, \ldots$, there must be one of lowest rank n_P such that the measure of the part of $C(E)$ that it contains is $> \frac{\alpha}{\lambda} m(d_{n_P})$. When d_{n_P} has been determined, we may suppose it to correspond to each of the points P, of E_α, which it contains. To any other point P', of E_α, not contained in d_{n_P}, there corresponds a similar mesh $d_{n_{P'}}$. An enumerable set of such meshes will contain all the points of E_α, and it is contained in Ω. The measure of the part of $C(E)$ contained in this enumerable set of meshes is $> \frac{\alpha}{\lambda} m(E_\alpha)$. But the measure of the set of meshes is $< m(E_\alpha) + \eta$, and therefore the measure of the part of $C(E)$ contained in the set of meshes is $< \eta$. Now η can be so chosen as to be $< \frac{\alpha}{\lambda} m(E_\alpha)$; hence the assumption that $m(E_\alpha) > 0$ leads to contradiction. It follows that $m(E_\alpha) = 0$. Taking for α the values in a decreasing sequence of numbers that converges to zero, the set of points of E at which the upper metric density of $C(E)$ is > 0 is the outer limiting set of the sets E_α, all of which have measure zero. Therefore those points of E at which the upper metric density of $C(E)$ is > 0 form a set of measure zero. With the exception of this set, at every point of E the metric density of $C(E)$ is zero.

It then follows that, at every point of E not belonging to a component of measure zero, the metric density of E is unity. Since E and $C(E)$ can be interchanged, we see that the metric density of E at all points of $C(E)$ is zero, except at points of a component of $C(E)$ of measure zero.

In case the set E is unbounded, we may consider the parts of it in cells of a sequence each of which contains the preceding one, and of which the measures increase indefinitely. Each of the cells may contain an exceptional set of measure zero, at which the metric density of the component of E in the cell either does not exist, or is not equal to unity. The outer limiting set of these exceptional sets has also the measure zero; hence the theorem holds for the unbounded set E.

The above theorem was first established, for the case of linear sets, by Lebesgue, who employed the theory of integration. Other proofs, independent of the theory of integration, have been given, for linear sets, by

Denjoy*, and by Lusin and Sierpiński†. The above proof is founded on the treatment of the subject by de la Vallée Poussin, who established the theorem for the case of sets of points in any number of dimensions.

THE RESOLUTION OF SETS OF POINTS IN ACCORDANCE WITH METRICAL PROPERTIES

141. It was shewn in § 91 that every set of points can be expressed as the sum of an enumerable (or finite) series of sets which have certain descriptive properties. We proceed to analyse a given measurable set into the sum of parts which have certain metrical properties.

It will be shewn that:

Any bounded measurable set can be expressed as the sum of an enumerable sequence of perfect sets, together with a set of which the measure is zero. Moreover, the perfect sets can be so chosen as to be all non-dense, and such that each of them is metrically dense in itself.

If S be a measurable set, it has been shewn in § 125 that it contains a perfect set H_1 such that $m(S) - m(H_1) < \epsilon/4$. If H_1 is not non-dense it has a non-dense perfect component G_1, such that $m(H_1) - m(G_1) < \epsilon/4$; therefore $m(S) - m(G_1) < \epsilon/2$. The perfect set G_1 has a perfect component (§ 138) L_1, of the same measure as G_1, and metrically dense in itself. We have now $S = L_1 + M_1$; where $m(S) - m(L_1) < \epsilon/2$; or $m(M_1) < \epsilon/2$. The set M_1 may be similarly resolved into $L_2 + M_2$; where L_2 is perfect, non-dense, and metrically dense in itself, and where $m(M_2) < \epsilon/2^2$. This procedure may be carried on successively, and terminates only if one of the sets M_n is either absent or has the measure zero. In case M_n exists for every value of n, its inner limiting set is M_ω, which has the measure zero, since $\lim\limits_{n\sim\infty} m(M_n) = 0$. We have therefore, in this case,

$$S = L_1 + L_2 + \ldots + L_n + \ldots + M_\omega,$$

where the series $L_1, L_2, \ldots$ of perfect sets, all non-dense and metrically dense in themselves, is either finite, or forms an unending series.

It will be observed that the set M_ω is a residual set, relatively to S, although it has the measure zero. In particular, the set of all the points of a cell can be resolved in this manner into a sum of sets; the residual S_ω being everywhere dense, and of cardinal number c, although its measure is zero.

---

* *Journal de Math.* (7), vol. I (1915), p. 132.

† *Rend. d. Circ. Mat. di Palermo*, vol. XLII (1917), p. 167. See also Sierpiński, *Fundamenta Mat.* vol. IV (1923), p. 167.

JORDAN'S MEASURE OF A SET OF POINTS

142. A definition has been employed by Jordan[*], and by Peano[†], of the measure of a set of points, which differs from that, developed by Borel and Lebesgue, which has been employed in the present Chapter. It is applicable to sets of points in space of any number of dimensions, and is of utility in the theory of quadratures.

Let a set G be in the interior of a closed cell (or interval) Δ; and let a system of nets with closed meshes be fitted on to Δ. Let $\Sigma_n^{(1)}$ be the sum of the measures of those meshes of D_n which are such that every point of each of them is an interior point of G, and let $\Sigma_n^{(2)}$ be the sum of those meshes of D_n which are such that each of them contains either an interior point of G or a point on the frontier of G and $C(G)$. It can be shewn that $\Sigma_n^{(1)}$ converges, as $n \sim \infty$, to a number S_1, and that $\Sigma_n^{(2)}$ converges to a number S_2; where S_1 and S_2 are independent of the particular system of nets fitted on to Δ

The number S_1 is called the interior extent of the set G, and S_2 is called the exterior extent of G When $S_1 = S_2$, the set is said to be measurable (J), and the number $S_1 \equiv S_2$ is said to be its measure (J). The exterior extent of a set is identical with its content, as defined by Harnack and Cantor (§ 118).

In accordance with this definition, any set which contains no interior points has its interior extent zero. For example, the interior extent of the irrational points of the linear interval $(0, 1)$ is zero, and its exterior extent is 1.

A set G consists, in general, of interior points forming a set I, and of points F_1, all of which belong to the frontier F. The interior extent of G is identical with the measure of the open set I, and the exterior extent of G is identical with the measure of the closed set $I + F$; the measures of these sets being defined as in § 122. It is then clear that the necessary and sufficient condition that a set G should be measurable (J) is that its frontier, which is a closed set, should have the measure zero. Accordingly, a set that is measurable (J) is also measurable in accordance with the definition here adopted (§ 128), but the converse does not hold.

THE SECTIONS OF A CLOSED SET

143. If G be any closed set of points in a rectangle $ABCD$, and through the points P, of AB, straight lines PP' are drawn perpendicular to AB, and if $f(P)$ denote the linear content of the linear component of G which is on the straight line PP', and which may be termed the section of G

---

[*] *Journal de Math.* (4), vol VIII (1892), p. 78; also *Cours d'Analyse,* vol. I, p. 28.

[†] *Applicazioni geom. del. calc. infinit.* (1887), p. 153.

by PP', then *the set of points P, on AB, which is such that $f(P) \geqq \sigma$, is a closed set; σ denoting any positive number.*

Let P_1 be a limiting point of the set; and if possible, let the linear content of that component of G which is on P_1P_1' be $< \sigma$; we can then determine a finite number of intervals $\delta_1, \delta_2, \ldots \delta_r$, on P_1P_1', whose sum is $> AD - \sigma$, and which are free in their interiors and at their ends from points of G. On each of these intervals δ we can describe a rectangle which contains no points of G within it or on its boundaries: this may be done on either side of P_1P_1'; for each point of δ can be enclosed in a rectangle free from points of G; and by the Heine-Borel theorem, a finite number of these rectangles, enclosing all the points of δ, exists. Take a point Q belonging to the set of points for which $f(Q) \geqq \sigma$, and let P_1Q be less than the breadth of all the rectangles described on the intervals $\{\delta\}$ on one side of P_1P_1'. On QQ' there is a finite number of intervals free from points of G, whose sum is $> AD - \sigma$, by the assumption as to P_1P_1'; hence the linear content of the component of G which is on QQ' must be $< \sigma$, which is contrary to the hypothesis. It follows that $f(P_1) \geqq \sigma$; hence the set of points on AB is closed.

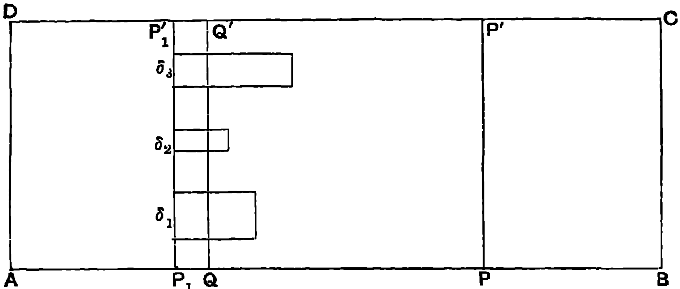

FIG. 2

It will now be shewn that, *for a closed set of points G, if for every position of P, on AB, the linear content of the section of G by PP' is $< \sigma$, then the content of G is $< \sigma . AB$.*

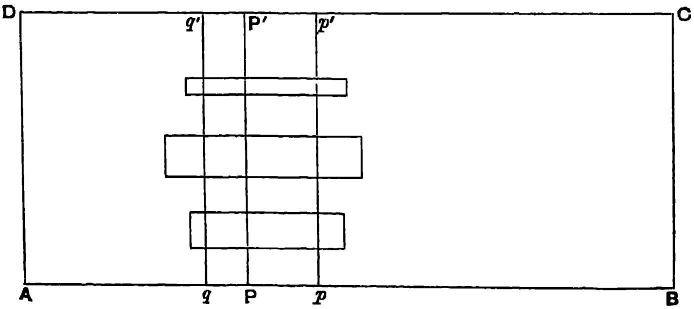

FIG. 3

Taking any point P, of AB; on PP', a finite number of intervals, whose sum is $> AD - \sigma$, can be found which are free from points of G; and on each of these intervals a rectangle can be drawn on each side of PP', containing no points of G in its interior or on its boundary. We can now draw two straight lines pp', qq', one on each side of P, so that each of them passes through the interiors of all the rectangles so described. We have now found an interval pq containing P, such that in $pqq'p'$ there is an area $> pq\,(AD - \sigma)$ free from points of G. Corresponding to each point P, of AB, such an interval pq can be found; and, in accordance with the Heine-Borel theorem, a finite number of these intervals can be selected, such that every point of AB is in the interior of one at least of them. The end-points of these intervals divide AB into a finite number of parts, such that, above any one part, of length a, there is an area $> a\,(AD - \sigma)$ free from points of G; and hence there is altogether an area $> AB\,(AD - \sigma)$ free from points of G. It follows therefore that the content of G is $< AB\,.\,\sigma$.

We shall now establish the following theorem, which is of importance in the theory of double integration:

If G be a closed set, and if the linear content of the set of points P, on AB, for which the linear content of the section of G by PP' is $\geq \sigma$, have the value zero, for every positive value of σ, then the set G is of zero content.

The points on AB, for which $f(P) \geq \sigma$, can be enclosed in a finite number of intervals whose sum is $< \epsilon$, where ϵ is an arbitrarily small number; and in each of the remaining parts of AB, the value of $f(P)$ is $< \sigma$; hence by the foregoing theorem the content of G is

$$< \sigma\,(AB - \epsilon) + \epsilon\,.\,AD;$$

and since this holds for arbitrarily small values of σ and ϵ, it follows that the content of G must be zero.

Conversely, it may be shewn that.

If G be a closed set, of plane content zero, the set of points P, on AB, for which the linear content of the section of G by PP' is $\geq \sigma$, has, for every positive value of σ, content zero. The linear measure of the set of points P for which the linear content of the section of G by PP' is greater than zero is zero.

Let I denote the linear content of the set (P) for which $f(P) \geq \sigma$; divide AB into n equal parts, and AD also into n equal parts, and through the end-points of these parts draw straight lines dividing the rectangle into equal parts each of area $\frac{1}{n^2}\,.\,AB\,.\,AD$. Then the sum of those parts of AB which contain points of the set (P) is always greater than I; and in each such part there is at least one point P, such that the sum of the parts of PP' which contain points of G is $> \sigma$. It follows that the sum of those rectangular portions which contain points of G is $> \sigma I$, however great n may be; and hence that the content of G is $\geq \sigma I$. Therefore it follows that G cannot have zero content unless I is zero.

The second part of the theorem is proved by giving to σ the values in a sequence $\{\sigma_n\}$ which converges steadily to zero. We have then only to consider the outer limiting set of the sequence of sets which correspond to the values of σ in the sequence $\{\sigma_n\}$.

EXAMPLES

1. Let a set of points (x, y) in the rectangle for which $0 \leq x \leq 1$, $0 \leq y \leq 1$, be defined as follows[*]:—The numbers x, y are expressed in the dyad scale, and only those values of x and y are taken which are expressed by terminating radix-fractions, the number of digits being the same for x as for y. If x' denotes a terminating radix-fraction, there is only a finite number of points (x', y') of the set on the straight line $x = x'$; similarly if y' denotes a terminating radix-fraction, there is only a finite number of points of the set on the straight line $y = y'$. The two-dimensional set is however everywhere dense; for, considering a straight line $y = x + a$, where a is a positive, or a negative, radix-fraction with a finite number of digits, we see that, corresponding to any number x expressed by a finite number of digits greater than the number of digits by which a is expressed, there is a point (x, y) on the straight line belonging to the set. The component of the set on the straight line $y = x + a$, being everywhere dense, and the values of a being everywhere dense in the interval $(-1, 1)$, it follows that the set is everywhere dense in the rectangle.

This example shews that an everywhere dense two-dimensional set may be linearly non-dense on each straight line belonging to two parallel sets. It also shews that a two-dimensional set may exist which is extended, but is unextended on straight lines belonging to either of two parallel sets.

2. Let a cross[†] formed by two pairs of straight lines parallel to the pairs of sides of a square be constructed, and so that the remainder of the square consists of four equal squares at the corners. Let the interior points of the cross be removed from the square, and then let a similar cross be removed from each of the remaining four squares. Proceeding in this manner, let the crosses be so chosen that the area of each square after the mth stage of the process is $ab^{-\frac{p}{2^m}}$ times the area of each square after the preceding stage. The sum of the areas of the squares which remain after the mth stage is

$$(4a)^m b^{-p\left(\frac{1}{2^m}+\frac{1}{2^{m-1}}+\cdot\ +\frac{1}{2}\right)} Q \equiv (4a)^m b^{-p\left(1-\frac{1}{2^m}\right)} Q,$$

where Q is the area of the original square. A non-dense closed set of points is defined as the points which remain when this process is carried on indefinitely. The limit of the sum of the crosses is that of

$$\left[1-(4a)^m b^{-p\left(1-\frac{1}{2^m}\right)}\right] Q;$$

and this is Q, or $< Q$, according as $a \leq \frac{1}{4}$; it follows that the closed set has content zero, if $a < \frac{1}{4}$; but if $a = \frac{1}{4}$, the content is $b^{-p} Q$.

---

[*] Pringsheim, *Sitzungsberichte d. Münch. Akad.* vol. XXIX (1899), p. 48.

[†] Veltmann, *Schlömilch's Zeitsch.* vol. XXVII (1882), pp. 178, 314.

CHAPTER IV

TRANSFINITE NUMBERS AND ORDER-TYPES

144. A preliminary account has been given, in Chapter II, of the theory of transfinite ordinal and cardinal numbers; it was shewn that the introduction of such numbers was suggested by the exigencies of the theory of linear sets of points, and that, in particular, the necessity for the use of transfinite ordinal numbers arises whenever a convergent sequence of points is transcended by adjoining to the points of the sequence their limiting point and any further points which it may be desirable to regard as belonging to the same set as the points of the sequence. The fundamental discovery of G. Cantor, that the rational points of an interval form an enumerable set, whereas the set of points of the continuum is unenumerable, by establishing the existence of a distinction between the characters of two infinite sets, suggests the development of a general theory of cardinal numbers of infinite aggregates. The procedure we adopted, of introducing the fundamental notions of transfinite ordinal and cardinal numbers in connection with the theory of sets of points, is in accord with the historical order in which the whole theory of transfinite numbers and order-types was developed. The account of the theory of transfinite numbers given in Chapter II is in general agreement with Cantor's earlier presentation* of his ideas; his later†, and more abstract, treatment of the subject is the one upon which the account given in the present Chapter is founded.

In order that the reader may be put into a position to form his own conclusions as to the validity of a scheme which must be regarded as still, to some extent at least, in the controversial stage, it has been thought best to postpone any discussion of the difficulties of the theory, until after the conclusion of the detailed account of the theory in its constructive aspect.

In the latter part of the Chapter, some critical remarks upon the logical basis of the theory will be made; these must necessarily be of an incomplete character, partly from considerations of space, and also because any complete criticism of such a scheme as Cantor's theory of transfinite numbers would involve the consideration of questions of an epistemological character, which for obvious reasons cannot be adequately dealt with in a work of a professedly mathematical complexion. Objections which may be urged against some parts of the theory will however be fully stated. Some consideration will also be given to the question, whether, and how far, the theory is indispensable as a logical basis of continuous Analysis.

---

* See his "Grundlagen einer allgemeinen Mannigfaltigkeitslehre," Leipzig, 1883, or *Math. Annalen*, vol. XXI (1883), p. 545; see also *Zeitschrift fur Phil. und phil. Kritik*, vols. LXXXVIII, XCI and XCII. Cantor's ideas were foreshadowed in a paradoxical form by Bolzano in his *Paradoxien des Unendlichen*, Leipzig, 1851, and although infinite numbers had been discussed by earlier writers, Bolzano is the only real predecessor of Cantor in this department of thought.

† This is contained in the two articles "Beitrage zur Begründung der transfiniten Mengenlehre," in the *Math. Annalen*, vol. XLVI (1895), and vol. XLIX (1897).

THE CARDINAL NUMBER OF AN AGGREGATE

145. *A collection* of definite distinct objects which is regarded as a single whole is called an aggregate.*

An aggregate may be denoted symbolically by a large letter M, the elements of the aggregate by small letters m; and the constitution of the aggregate may be denoted by the equation $M = \{m\}$.

The consideration of questions which arise in connection with this definition, as to the mode in which the objects of the aggregate must be specified in order that the aggregate may be adequately defined, and as regards the conditions, if any, which must be satisfied in order that a collection may be regarded as a whole, or aggregate, of such a character that it can be an object of mathematical thought, will be postponed. For the present, it is sufficient to remark that an adequate definition of any particular aggregate, which is not necessarily finite, must contain, as a minimum, a set of rules or specifications by means of which it is *theoretically* determinate, in respect of any object whatever, whether such object does, or does not, belong to the aggregate. The set of prime numbers, for example, is regarded as an aggregate although, when a particular number is presented to us, we may be *practically* unable to decide whether that number is prime or not. In this case, however, a finite number of processes will suffice to decide the question. If however, we take the case of the algebraical numbers, the state of things is different; for we are not in possession of any general method which enables us to decide whether a given number is algebraic or not. Nevertheless, the question being regarded as having a logically determinate answer, the algebraical numbers are regarded as forming an aggregate, in the sense here employed.

An aggregate does not depend, for its validity as a mathematical entity, upon the possibility of producing all its members, successively or otherwise, but upon the sufficiency of the rules by which its elements are to be distinguished, as belonging to it, in that particular kind of objects to which they belong; that is, upon the sufficiency, in this direction, of its definition of membership.

---

* This definition is given by Cantor, *Math. Annalen*, vol. XLVI (1895), p. 481, as follows:—"Unter einer 'Menge' verstehen wir jede Zusammenfassung M von bestimmten wohl unterschiedenen Objecten m unserer Anschauung oder unseres Denkens (welche die 'Elemente' von M genannt werden) zu einem Ganzen."

Two aggregates M, N are said to be *equivalent* to one another when they are such that a law of correspondence can be established between the elements of one aggregate, and those of the other, so that, to each element of one of the aggregates, there corresponds one, and only one, element of the other aggregate.

This relation of equivalence between two aggregates M and N may be expressed symbolically by $M \sim N$, or $N \sim M$.

It is clear that, if each of two aggregates is equivalent to a third, the two aggregates are equivalent to one another.

Aggregates which are equivalent to one another are said to have the same power, or cardinal number.

A cardinal number is accordingly characteristic of a class of equivalent aggregates.

The question whether two defined aggregates have, or have not, the same cardinal number, is thus equivalent to the question whether it is, or is not, possible to establish a systematic (1, 1) correspondence between the elements of the two aggregates, in accordance with the above definition of equivalence.

A particular aggregate can ordinarily be shewn to be equivalent to itself. The law of correspondence between an element and another element which can be set up is in general of a character which admits of a certain arbitrariness. The cardinal number is accordingly regarded as independent of the notion of order in the aggregate.

The power, or cardinal number, of an aggregate M has been defined by Cantor as the concept which is obtained by abstraction when the nature of the elements of M, and the order in which they are given, are entirely disregarded.

Cantor regards the fact that equivalent aggregates have the same cardinal number as a deduction from this definition.

The cardinal number of M is a characteristic of M which may be denoted by $\overline{\overline{M}}$, to indicate that both the order of the elements, and their precise individual nature, are irrelevant as regards the cardinal number.

The relation of equivalence $M \sim N$, between two aggregates, implies the equality $\overline{\overline{M}} = \overline{\overline{N}}$; and this equation expresses the necessary and sufficient condition for the equivalence of M and N.

Since Cantor regards the cardinal number of M as independent of the precise nature of the elements of M, we may, in accordance with this view, substitute for each element the number unity. We have thus a new aggregate which is a collection of elements each of which is the number 1, and is equivalent to M; and this new aggregate is regarded by Cantor as a symbolical representation of the cardinal number $\overline{\overline{M}}$.

THE RELATIVE ORDER OF CARDINAL NUMBERS

146. Every aggregate M_1, which is such that all its elements are also elements of M, is called a *part*, or *sub-aggregate*, of M.

If M_2 is a part of M_1, and M_1 is a part of M, then M_2 is a part of M.

A finite aggregate cannot be equivalent to any of its sub-aggregates; but as will be seen in detail further on, an infinite aggregate always possesses sub-aggregates which are equivalent to itself. This is the characteristic distinction between finite and infinite aggregates, and has in fact been employed by Dedekind and others to define an infinite aggregate as one which is equivalent to one of its parts.

If two aggregates M, N, with the cardinal numbers $\alpha \equiv \overline{\overline{M}}$, $\beta \equiv \overline{\overline{N}}$, are such that, (1), there exists no part of M which is equivalent to N, and (2), there exists a part N_1, of N, which is equivalent to M, it is clear that the corresponding conditions are satisfied for any two aggregates which are equivalent to M, N respectively; and thus the two conditions characterize a relation between the cardinal numbers α, β of the two aggregates. When the above conditions are satisfied we say that α is *less* than β, and that β is *greater* than α; which is expressed symbolically by $\alpha < \beta$, $\beta > \alpha$. This is the definition of inequality for two cardinal numbers, and of the relations greater and less, in the purely ordinal sense in which they are here used.

The condition contained in the definition is inconsistent with the relation of equality between α and β being satisfied. For if $\alpha \equiv \beta$, then $M \sim N$; hence since $N_1 \sim M$, we have $N_1 \sim N$: therefore, since $M \sim N$, there must be a part of M, say M_1, such that $M_1 \sim M$, which would involve $M_1 \sim N$; but this is contrary to one of the conditions contained in the definition of inequality.

It is easily seen that if $\alpha < \beta$, and $\beta < \gamma$, then $\alpha < \gamma$.

147. It has been seen that the three relations $\alpha = \beta$, $\alpha < \beta$, $\alpha > \beta$ are mutually exclusive; but the question arises whether any two cardinal numbers α, β whatever must satisfy one of these relations. An affirmative answer to this question would be required before it could be maintained that all cardinal numbers can be regarded as being alike capable of having relative rank assigned to them, in a single ordered aggregate.

Two aggregates M, N, of which we may denote parts by M_1, N_1, must satisfy one, and only one, of the following four conditions:

(1). M, N have parts M_1, N_1, such that $M_1 \sim N$, and $N_1 \sim M$.

(2). M has a part M_1, such that $M_1 \sim N$; but no part of N exists which is equivalent to M.

(3). There is no M_1 which is equivalent to N; but there is an N_1 which is equivalent to M.

(4). There exists no M_1 equivalent to N; and also no N_1 equivalent to M.

It will be proved that, if the condition (1) is satisfied, then $\overline{\overline{M}} = \overline{\overline{N}}$. The condition (2) expresses the relation defined as $\overline{\overline{M}} > \overline{\overline{N}}$. The condition (3) expresses the relation defined as $\overline{\overline{M}} < \overline{\overline{N}}$.

In case M is a part of N, M is itself equivalent to a part of N, and thus either (1) or (2) is satisfied; and therefore $\overline{\overline{M}} \leq \overline{\overline{N}}$.

It has not been proved that the relation (4) is an impossible one: except that, in the case of finite aggregates, it may be easily seen that it involves $\overline{\overline{M}} = \overline{\overline{N}}$. Until this point is cleared up, it cannot be maintained as an established fact that the cardinal numbers α, β of any two aggregates whatever satisfy one of the three relations $\alpha = \beta$, $\alpha > \beta$, $\alpha < \beta$.

Two aggregates which are such that their cardinal numbers α, β stand to one another in one of the relations $\alpha = \beta$, $\alpha > \beta$, or $\alpha < \beta$, may be said to be *comparable* with one another. Otherwise they are *incomparable* with one another.

THE ADDITION AND MULTIPLICATION OF CARDINAL NUMBERS

148. If M, N are two aggregates which have no element in common, then the aggregate which has for its elements all those of M and all those of N is called the sum of the two aggregates M, N, and may be denoted by (M, N). A similar definition applies to the case of the sum of any number of aggregates, no two of which have an element in common.

If M', N' are two other aggregates with no element in common, such that $M \sim M'$, $N \sim N'$, it is clear that $(M, N) \sim (M', N')$; and thus the cardinal number of (M, N) depends only on those of M and N.

If $\overline{\overline{M}} = \alpha$, $\overline{\overline{N}} = \beta$, *we define the result of the operation of addition of* α *and* β *to be* $(\overline{\overline{M, N}})$.

From the independence of cardinal numbers of the order of elements, we deduce

$$\alpha + \beta = \beta + \alpha, \quad \alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma;$$

thus the operation of addition of cardinal numbers obeys the commutative and associative laws.

149. If an element m, of M, be associated with an element n, of N, so as to form a new element (m, n), the aggregate of all possible elements which can be formed in this way is called the product of M and N, and may be denoted by $(M . N)$.

If $M \sim M'$, $N \sim N'$, it is clear that, to each element (m, n), of $(M . N)$, there is a corresponding element of $(M' . N')$, hence $(M . N) \sim (M' . N')$; and thus $(\overline{\overline{M . N}})$ depends only on $\overline{\overline{M}}$ and $\overline{\overline{N}}$.

The cardinal number of the product-aggregate $(M.N)$ *defines the product of the cardinal numbers of* M *and* N.

The product of $\overline{\overline{M}}$ *and* $\overline{\overline{N}}$ *may also be defined as the cardinal number of the aggregate which is obtained by substituting for each element of* N *an aggregate which is equivalent to* M.

It is seen on reflection that this definition is equivalent to the first one.

Since, as can be shewn from the definition,

$$(M.N) \sim (N.M),\ (M.(N.R)) \sim ((M.N).R)$$

and

$$(M.(N, R)) \sim ((M.N), (M.R)),$$

we see that cardinal numbers satisfy the relations

$$\alpha.\beta = \beta.\alpha;\ \alpha.(\beta.\gamma) = (\alpha.\beta)\gamma;\ \alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma.$$

It has thus been shewn that the multiplication of cardinal numbers obeys the commutative, associative, and distributive laws.

The definition of multiplication may be extended* to the case in which the number of factors is not necessarily finite. Let us consider a class of aggregates M, where the class contains either a finite, or an infinite, number of aggregates, and suppose no two of the aggregates have an element in common. Let a new object consist of an association of elements of the aggregates in the given class, one element belonging to each of those aggregates. All such objects may be regarded as the elements of a new aggregate. This new aggregate is said to be the product-aggregate of the given class of aggregates, and its cardinal number defines the product of the cardinal numbers of all the aggregates of the given class.

CARDINAL NUMBERS AS EXPONENTS

150. If we have two finite aggregates M, N, containing x and y elements respectively, we may suppose that, to each of the y elements of N, one element of M is made to correspond, so that the same element of M may be used any number of times; any such particular correspondence we call a distribution of N upon M. The total number of ways of distributing N upon M is x^y. To put the matter in a concrete form, the total number of ways of distributing y things among x persons, where any number of the y things may be given to one person, is x^y; any particular mode of distribution is what we have called a mode of distributing the aggregate of y things upon the aggregate of x persons.

The definition of distributing an aggregate N upon an aggregate M is immediately extensible to the case of infinite aggregates. As before, the distribution denotes any system by which, to each element of N, is made to correspond a particular element of M, the same element of M being employed any number of times, or not at all. Denoting by N/M each particular mode of distributing N upon M, we thus form the new aggregate (N/M) which contains as its elements all such distributions.

---

* See Whitehead, *American Journal of Math.* vol. xxiv (1902), p. 367, where the theory of cardinal numbers is treated by the Peano-Russell symbolical method.

It is seen at once that, if $M \sim M'$, $N \sim N'$, then $(N/M) \sim (N'/M')$. Thus the cardinal number of (N/M) depends only on the cardinal numbers of M and N.

The cardinal number of the aggregate (N/M), each element of which is a distribution of N on M, and in which every possible mode of such distribution occurs as an element, is denoted by the symbol α^β, where $\alpha \equiv \overline{\overline{M}}$, $\beta \equiv \overline{\overline{N}}$; thus $\alpha^\beta \equiv (\overline{\overline{N/M}})$.

It is easy to shew that

$$((N/M).(R/M)) \sim ((N, R)/M)$$
$$((R/M).(R/N)) \sim (R/(M.N))$$
$$(R/(N/M)) \sim ((R.N)/M).$$

Hence if $\overline{\overline{M}} = \alpha$, $\overline{\overline{N}} = \beta$, $\overline{\overline{R}} = \gamma$, we see that, in accordance with the above definition of exponentials,

$$\alpha^\beta . \alpha^\gamma = \alpha^{\beta+\gamma}, \quad \alpha^\gamma . \beta^\gamma = (\alpha . \beta)^\gamma, \quad (\alpha^\beta)^\gamma = \alpha^{\beta \cdot \gamma};$$

and thus the same laws hold as for exponents in which only finite cardinal numbers are involved.

THE SMALLEST TRANSFINITE CARDINAL NUMBER

151. The cardinal number of the aggregate of all the finite integers 1, 2, 3, ... n, ... is called Alef-zero, and is denoted by $\aleph_0$; thus $\aleph_0 = \{\overline{\overline{n}}\}$. The number $\aleph_0$ is identical with the number which has previously been denoted by a.

If we add to $\{n\}$ a new element e, we obtain the sum-aggregate $(\{n\}, e)$, and this is equivalent to $\{n\}$, for we may make e in the first of these aggregates correspond to 1 in the second, and in general n to $n+1$; and thus $(\{n\}, e) \sim \{n\}$. From this, we obtain $\aleph_0 + 1 = \aleph_0$, a relation which differentiates $\aleph_0$ from all the finite cardinal numbers.

The cardinal number $\aleph_0$ is greater than all the finite cardinal numbers, and no transfinite cardinal number exists that is less than $\aleph_0$.

Since the finite aggregate $(1, 2, 3, \dots k)$ is a part of $\{n\}$, but no part of the finite aggregate is equivalent to $\{n\}$, we have, by the definition of inequality, $\aleph_0 > k$.

Let us suppose, if possible, that $\alpha \equiv \{\overline{\overline{M}}\} < \aleph_0$; then there is no part of $\{M\}$ equivalent to $\{n\}$, but there is a part of $\{n\}$ equivalent to $\{M\}$. Now every part of $\{n\}$ is either finite, or else is equivalent to $\{n\}$. For, consider a part N_1, of $\{n\}$; this must contain some number $\bar{n}$; either $\bar{n}$ is the number of lowest rank in N_1, or there are in N_1 numbers of lower rank than $\bar{n}$. In either case N_1 contains a part N_2 which is a part of the finite aggregate $1, 2, 3, \dots \bar{n}$. In accordance with the definition in § 2, N_2 has a number n_1, of lower rank than all the other numbers in $1, 2, 3, \dots \bar{n}$; this number n_1 has lower rank than all the other numbers in N_1. Removing n_1 from N_1, we now have a part N_2, of $\{n\}$, and it may be proved as before that there exists a number n_2, of lower rank than all the other numbers in N_2. Proceeding in this manner, we obtain a set $n_1, n_2, n_3, \dots$ of numbers all belonging to N_1, each one of which is of lower rank than the next. If this sequence terminates at some number n_p, N_1 is a finite aggregate; if it does not terminate, it is a sequence equivalent to $\{n\}$. Since every element of N_1 must occur as one of the numbers $\{n_p\}$, it thus follows that N_1 is equivalent to $\{n\}$. It now follows that, either $\{M\}$ is a finite aggregate, in which case its cardinal number is $< \aleph_0$, or else $\{M\}$ is equivalent to $\{n\}$, in which case its cardinal number is $\aleph_0$, contrary to the hypothesis that it is $< \aleph_0$. It has thus been shewn that the cardinal number of any infinite aggregate is either $\geqq \aleph_0$, or else that it is incomparable with $\aleph_0$.

The following more stringent theorem was proved by Cantor:

The cardinal number $\aleph_0$ is less than any other transfinite cardinal number.

The proof of this theorem, unlike that of the foregoing one, requires the employment of an infinite process of choosing particular elements from an aggregate. It is equivalent to the theorem that *every non-finite aggregate contains a part which is equivalent to $\{n\}$, the aggregate of finite numbers, that is, a part whose cardinal number is $\aleph_0$.*

If M be any non-finite aggregate, the process of choosing a first, second, ... nth element from the aggregate can be continued indefinitely without limit. Thus M contains a part that is equivalent to $\{n\}$, and therefore (§ 147) its cardinal number is $\geqq \aleph_0$. Thus every cardinal number that is not finite, and is different from $\aleph_0$, is $> \aleph_0$.

152. It has been shewn that $\aleph_0 + 1 = \aleph_0$: a similar proof would shew that $\aleph_0 + n = \aleph_0$, where n is any finite integer.

In accordance with the definition of addition, $\aleph_0 + \aleph_0$ is the cardinal number of the aggregate $(1, 3, 5, \dots 2, 4, 6, \dots)$, for $\aleph_0$ is the cardinal number of each of the aggregates $(1, 3, 5, \dots)$ $(2, 4, 6, \dots)$; hence, since the cardinal number of $(1, 3, 5, \dots 2, 4, 6, \dots)$ is the same as that of $\{n\}$, we have $\aleph_0 + \aleph_0 = \aleph_0$, which we may write as

$$\aleph_0 . 2 = 2 . \aleph_0 = \aleph_0 .$$

From this relation, by repeated addition of $\aleph_0$ to both sides of the identity, we find

$$\aleph_0 . n = n . \aleph_0 = \aleph_0 .$$

In order to express the product $\aleph_0 . \aleph_0$, we form the aggregate $\{(n, n')\}$ of which the elements (n, n') consist of every pair of finite cardinal numbers. Let $n + n' = s$, then s has the values $2, 3, 4, \ldots$; and for any fixed value of s the numbers n, n' have a definite number of sets of values. Let $s = 2$: we then have one element $(1, 1)$: let $s = 3$, we then have two elements $(1, 2)$, $(2, 1)$: for $s = 4$, we have $(1, 3)$, $(2, 2)$, $(3, 1)$, and so on. The elements of $\{(n, n')\}$ may thus be arranged in order so that the element (n, n') is at the pth place, where

$$p = n + \tfrac{1}{2}(n + n' - 1)(n - n' - 2);$$

thus the aggregate $\{(n, n')\}$ is equivalent to $\{p\}$, which has the cardinal number $\aleph_0$.

It has now been proved that $\aleph_0 . \aleph_0 = \aleph_0$, or $\aleph_0^2 = \aleph_0$, and from this the theorem $\aleph_0^n = \aleph_0$ follows, by repeated multiplication by $\aleph_0$.

The theorems $n . \aleph_0 = \aleph_0$, $\aleph_0^2 = \aleph_0$ express in a symbolical form the results which have been proved in § 58, that a finite, or an enumerably infinite, set of enumerable aggregates makes an enumerable aggregate.

THE EQUIVALENCE THEOREM

153. The proof referred to in § 147 will now be given, that, if M, N are any two aggregates such that M contains a part M_1 which is equivalent to N, and N contains a part N_1 equivalent to M, then $\overline{\overline{M}} = \overline{\overline{N}}$. This theorem, which may be called the equivalence theorem, was first proved by Schröder* and independently by Bernstein†; but the form in which the proof is here given is due to Zermelo‡.

Lemma I. If a cardinal number α remains unaltered by the addition of any one of the enumerable set of cardinal numbers $p_1, p_2, \ldots p_n, \ldots$, it remains unaltered if all these cardinal numbers p are added to it at once.

Let $M, P_1, P_2, \ldots P_n, \ldots$ be aggregates of which the cardinal numbers are $\alpha, p_1, p_2, \ldots p_n, \ldots$; and such that $P_1, P_2, \ldots P_n, \ldots$ are all parts of M. We have then,

$$M = (P_1, M_1) = (P_2, M_2) = \ldots = (P_n, M_n) = \ldots;$$

where $M_1, M_2, \ldots$ are all parts of M, and in virtue of the hypothesis made in the statement of the theorem,

$$\overline{\overline{M}} = \overline{\overline{M}}_1 = \overline{\overline{M}}_2 = \ldots = \overline{\overline{M}}_n = \ldots$$

for $p_n + \overline{\overline{M}} = \overline{\overline{M}} = p_n + \overline{\overline{M}}_n$.

---

* See *Jahresbericht d. deutsch. math. Vereinigung*, vol. v (1896), p. 81; also *Nova Acta Leop.* vol. LXXI (1898), p. 303.

† See Borel's *Leçons sur la théorie des fonctions*, p. 103.

‡ *Göttinger Nachrichten*, 1901, p. 34, "Ueber die Addition transfiniter Cardinalzahlen." For remarks upon these proofs see Korselt, *Math. Annalen*, vol. LXX (1911), p. 294.

We may denote the (1, 1) correspondence which can be set up (see § 145) between M and M_n, by $M_n = \phi_n M$; and this for every n. Now it is clear that this relation of correspondence is such that

$$\phi M = (\phi P_1, \phi M_1) = (\phi P_2, \phi M_2) = \dots .$$

Hence $M = (P_1, M_1)$,

$$M_1 = \phi_1 M = (\phi_1 P_2, \phi_1 M_2) = (P_2', M_2'),$$

where P_2', M_2' are those aggregates which correspond to P_2, M_2, respectively, in the correspondence denoted by ϕ_1.

Also, with a similar notation,

$$M_2' = \phi_1\phi_2 M = (\phi_1\phi_2 P_3, \phi_1\phi_2 M_3) = (P_3', M_3'),$$

. . ..,

$$M'_{r-1} = \phi_1\phi_2 \dots \phi_{r-1} M = (\phi_1\phi_2 \dots \phi_{r-1} P_r, \phi_1\phi_2 \dots \phi_{r-1} M_r) = (P_r', M_r').$$

From these results we deduce

$$M = (P_1, P_2', P_3', \dots P_r', M_r');$$

and no two of the parts $P_1, P_2', P_3', \dots P_r'$, of M, have elements in common. This process of division of M can be continued indefinitely; and we then have $M = (P_1, P_2', P_3', \dots M_\omega')$,

where P_r', for every r, is included, and M_ω' consists of those elements which belong to M_r' for every value of r. From this we see that

$$a = p_1 + p_2 + p_3 + \dots + a';$$

where a' is the cardinal number of M_ω'.

Let us now consider the special case of the lemma which arises when $p_1, p_2, \dots$ are all equal, say to p In this case, we see that, from the hypothesis $a = p + a$, the result $a = \aleph_0 p + a'$ follows, where $\aleph_0$ denotes the cardinal number of the series of finite integers.

Now since $\aleph_0 = 2\aleph_0$, we have $\aleph_0 p + a' = 2\aleph_0 p + a' = \aleph_0 p + a$; it has thus been shewn that, if $a = a + p$, then $a = a + \aleph_0 p$.

Returning to the general case, we have

$$a = a + \aleph_0 p_1 = a + \aleph_0 p_2 = \dots;$$

it now follows that $a = \aleph_0 p_1 + \aleph_0 p_2 + \dots + a''$,

where a'' is the value which a' takes when $\aleph_0 p_1, \aleph_0 p_2, \dots$ are substituted for $p_1, p_2, \dots$.

Hence we have

$$a = 2\aleph_0 (p_1 + p_2 + \dots) + a'' = a + \aleph_0 (p_1 + p_2 + \dots)$$
$$= (\aleph_0 + 1)(p_1 + p_2 + \dots) + a'' = a + p_1 + p_2 + \dots :$$

and therefore the lemma has been established.

Lemma II. If the sum of two cardinal numbers p and q, when added to a, leaves a unaltered, then a is unaltered by the addition of either p or q.

For if $$\alpha = \alpha + p + q,$$
we have seen that $$\alpha = \alpha + \aleph_0 (p + q);$$
hence $$\alpha = \alpha + (\aleph_0 + 1) p + \aleph_0 q,$$
and also $$\alpha = \alpha + \aleph_0 p + (\aleph_0 + 1) q;$$
from these equalities we have
$$\alpha = \alpha + p, \text{ and } \alpha = \alpha + q.$$

We are now in a position to prove the equivalence theorem. If
$$\alpha = \beta + p, \text{ and } \beta = \alpha + q,$$
we have $$\alpha = \alpha + p + q,$$
and hence, by Lemma II,
$$\alpha = \alpha + p = \alpha + q = \beta,$$
therefore, if M has a part equivalent to N, and N has a part equivalent to M, it follows that $\overline{\overline{M}} = \alpha + p = \alpha + q = \overline{\overline{N}}$; where α is the cardinal number of the part of M that is equivalent to N, and β is that of the part of N that is equivalent to M.

In case M is itself equivalent to N, which will in particular happen if M is a part of N, we have $p \equiv 0$, and the theorem holds good.

In case the condition $\alpha = \beta + p$ holds, but there is no corresponding condition $\beta = \alpha + q$, we have in accordance with the definition in § 147, $\alpha > \beta$. It follows that *the sum of two or more cardinal numbers is greater than, or equal to, any one of the cardinal numbers.*

The following theorem may be established:

If the cardinal number α is unaltered by the addition of p, and if $\beta \geqq \alpha$, the cardinal number β is unaltered by the addition of q, where $q \leqq p$.

For let $\beta = \alpha + \gamma$, $p = q + r$; then, from $\alpha = \alpha + p = \alpha + q + r$, we deduce that $\alpha = \alpha + q$. It then follows that
$$\beta = \alpha + \gamma = \alpha + q + \gamma = \alpha + r + \gamma = \beta + q = \beta + r.$$

154. A proof has been given by Cantor* that, *if an aggregate exists of which the cardinal number is α, then an aggregate exists of which the cardinal number is greater than α.*

The proof is a generalization of the second proof, given in § 60, that the cardinal number c, of the continuum, is greater than that of the rational numbers. The proof may be put into the following form:

Suppose $M = \{m\}$ to be an aggregate, of cardinal number α; this aggregate M may be supposed to be simply ordered (see § 157) in any manner. In M let each element m be replaced either by A or by B, where A, B are two given objects; then M is replaced by a similar aggregate (see § 157), in which each element is either A or B. An infinity of such aggregates will be obtained, differing from one another in respect of whether A or B has been put in the place of each element of M; denoting the aggregate of all such possible aggregates M_{AB}, by $\{M_{AB}\}$, it will be shewn that the cardinal number of $\{M_{AB}\}$ is greater than that of M. In the first place, it can be seen that the cardinal number of $\{M_{AB}\}$ is equal to, or greater than, that of $\{m\}$; for, taking any one element m_0, of $\{m\}$, replace it by A, and all the other elements by B; we have then an element of $\{M_{AB}\}$, and there is such an element corresponding to each element m_0, of $\{m\}$; thus those elements of $\{M_{AB}\}$ in which there is only one A, form an aggregate of cardinal number equal to that of $\{m\}$. Next, let us assume that, if possible, all the elements of $\{M_{AB}\}$ are placed in (1, 1) correspondence with those of $\{m\}$; it will then be shewn that an M_{AB} can always be defined which is not included in the correspondence. Each $M^0{}_{AB}$, in $\{M_{AB}\}$, now corresponds to a definite m_0, in $\{m\}$; form a new aggregate M'_{AB} in the following manner:—For each element $M^0{}_{AB}$, in $\{M_{AB}\}$, in which A takes the place of m_0, in $\{m\}$, substitute B; and for each element $M^0{}_{AB}$, in $\{M_{AB}\}$, in which B takes the place of m_0, in $\{m\}$, substitute A; in this manner we form an aggregate M'_{AB} in which each element is either A or B, which is similar to $\{m\}$, and which is not identical with any M_{AB} that occurs in the correspondence between $\{M_{AB}\}$ and $\{m\}$. It has thus been shewn that the cardinal number of the aggregate of all the M_{AB} is greater than that of M. If M is the aggregate $a_1, a_2, a_3, \ldots a_n, \ldots$ which is similar to the aggregate of integral numbers, and if, for A and B, we take 0 and 1 respectively, then the aggregate $\{M_{01}\}$ may be interpreted as the aggregate of all the rational and irrational numbers between 0 and 1, in the binary scale; and this aggregate is thus shewn to be unenumerable.

---

* See *Jahresbericht d. deutsch. math. Vereinigung*, 1897.

Instead of replacing the elements of $\{m\}$ by two letters A, B, we might have taken any finite number of letters, without altering the principle of the proof. In § 60, the ten digits 0, 1, ... 9 were taken instead of A and B. It will be observed that, even if $\{m\}$ is normally ordered (see § 165), the new aggregate $\{M\}$ is not given as a normally ordered aggregate; and in default of proof it cannot be assumed that it is capable of being arranged in normal order.

To replace all the elements of an aggregate either by A, or by B, is equivalent to taking a part* of the given aggregate. The theorem has thus been established that, *the cardinal number of the aggregate, each element of which is a part of a given aggregate, is greater than the cardinal number of the given aggregate, all possible parts being contained in the new aggregate.*

155. An important question relating to cardinal numbers in general arises as regards the sum, $\alpha + \beta$, of two transfinite cardinal numbers.

---

* See Borel, *Leçons sur la théorie des fonctions*, p. 108.

If $\alpha + \beta = \gamma$, it may be the case that γ is greater than both the numbers α, β, or that it is equal to one of them; it being assumed that α, β are comparable cardinal numbers. Whether the first case can arise, or not, has not been definitely settled*, even in the case in which $\alpha = \beta$; thus it is not known whether 2α is necessarily equal to α, as is the case when $\alpha = c$, or when $\alpha = \aleph_0$.

In case $\alpha = 2\alpha = \alpha + \alpha$, it then follows, by applying Lemma I of § 153, that $\alpha = \alpha + \aleph_0\alpha = (\aleph_0 + 1)\,\alpha$; and thus that $\alpha = \aleph_0\alpha$.

The special case in which α is the cardinal number of a set of points in space of any number of dimensions has been referred to in § 90, where the hypothesis has been alluded to that, if x be the cardinal number of such set, then $ax = x$. That this holds good, involves the assumption that $2x = x$; which is equivalent to the assertion that, if there exists a set, of cardinal number x, in each of two non-overlapping cells, the cardinal number of the combined set is also x

DIVISION OF CARDINAL NUMBERS BY FINITE NUMBERS

156. If two aggregates have the same cardinal number, and if each of the two aggregates be divided into the same finite number n, of parts, such that the n parts of the first aggregate all have the same cardinal number, and also the n parts of the second all have the same cardinal number, then it can be proved that the cardinal number of one of the parts of the first aggregate is the same as that of one of the parts of the second aggregate. Symbolically, the theorem may be stated in the form:— if α, β are cardinal numbers such that $n\alpha = n\beta$, then $\alpha = \beta$.

This theorem has been proved by Bernstein†. It will be sufficient to give the detailed proof in the case $n = 2$, as the proof in the general case is obtained by generalization of that employed in the particular case

Since an aggregate is equivalent to itself, any special mode of exhibiting such equivalence, by which each element is made to correspond to a definite other element, may be called a transformation of the system into itself. As regards all such possible transformations the following propositions may be seen to hold:

(1). The transformations of an aggregate M into itself form a *group* ϕ_M.

(2). Let $1, \chi_1, \chi_2, \chi_3, \ldots$ denote a sequence of transformations of M into itself, 1 denoting the identical transformation, and let this sequence form a group which is necessarily a sub-group of ϕ_M; then the condition that the sequence forms a group is that, corresponding to any two integers

---

* See Schoenflies, *Die Entwickelung...*, vol. II, p. 9; also Jourdain, *Phil. Mag.* (6), vol. VI (1903), p. 323.

† Inaugural Dissertation, "Untersuchungen aus der Mengenlehre," Halle, 1901. This is reproduced in *Math. Annalen*, vol. LXI (1905), p. 117.

m, n, there is a third r, such that $\chi_m \chi_n = \chi_r$. Further, let us suppose that, to every χ_n there corresponds a definite χ'_n, such that $\chi_n \chi'_n = 1$. If m be an element of M, such that $m \neq \chi_n(m)$, for $n = 1, 2, 3, \ldots$, then

$$\chi_n(m) \neq \chi_{n'}(m),$$

where n and n' are any unequal integers.

(3). If m and m' are any two distinct elements of M, and if $m \neq \chi_n(m')$, for $n = 1, 2, 3, \ldots$, then $\chi_n(m) \neq \chi_{n'}(m')$. For if $\chi_n(m) = \chi_{n'}(m')$, we should deduce that $m = \chi'\chi_n(m) = \chi'\chi_{n'}(m') = \chi_{n''}(m')$, which is contrary to the hypothesis made.

(4). If $T_1, T_2, \ldots$ are parts of M, such that each T has no element in common with another T, we may say that the T's form a system of separate parts of M.

If $T = \{t\}$ is a part of M, and if $t \neq \chi_n(t')$, for $n = 1, 2, 3, \ldots$, then the equivalent aggregates $T, \chi_1(T), \chi_2(T), \ldots$ form a system of separate parts of M.

(5). If T is a part of M which satisfies the condition stated in (4), then $\overline{\overline{M}} = \overline{\overline{M}} + \overline{\overline{T}}$.

For $T, \chi_1(T), \chi_2(T), \ldots$ are all parts of M having the cardinal number $\overline{\overline{T}}$; and if R is the part of M which remains when all these separate parts are removed, we have

$$\overline{\overline{M}} = \overline{\overline{R}} + \aleph_0 . \overline{\overline{T}};$$

hence

$$\overline{\overline{M}} + \overline{\overline{T}} = \overline{\overline{R}} + (\aleph_0 + 1)\overline{\overline{T}}$$
$$= \overline{\overline{R}} + \aleph_0 . \overline{\overline{T}} = \overline{\overline{M}}.$$

After these remarks we can proceed to the proof of the theorem:—Let

$$(a), \qquad \overline{\overline{M}} = \overline{\overline{x}}_1 + \overline{\overline{x}}_2 = \overline{\overline{x}}_3 + \overline{\overline{x}}_4,$$
$$(b), \qquad \overline{\overline{x}}_1 = \overline{\overline{x}}_2,$$
$$(c), \qquad \overline{\overline{x}}_3 = \overline{\overline{x}}_4;$$

then it is required to shew that $\overline{\overline{x}}_1 = \overline{\overline{x}}_3$; which involves $\overline{\overline{x}}_2 = \overline{\overline{x}}_4$.

The three equations (a), (b), (c) may be regarded as denoting that there are three reversible transformations of the aggregate M into itself, which may be denoted by ϕ_a, ϕ_b, ϕ_c respectively; the reversibility of these transformations is expressed by $\phi_a^2 = \phi_b^2 = \phi_c^2 = 1$.

The transformation ϕ_a involves $x_1 = (x_{13}, x_{14})$, where x_{13} are those elements of x_1 which are transformed into elements of x_3, and x_{14} those which are transformed into elements of x_4; on the whole we have

$$(6). \qquad \begin{cases} x_1 = (x_{13}, x_{14}), \\ x_2 = (x_{23}, x_{24}), \\ x_3 = (x_{31}, x_{32}), \\ x_4 = (x_{41}, x_{42}), \end{cases} \qquad \text{where } \overline{\overline{x}}_{ik} = \overline{\overline{x}}_{ki}.$$

If T_1 is any part of x_1, and T_2 an equivalent part of x_2, we may denote which belong to T_1 with those of x_2 which belong to T_2; we have then a similar set of equations to (6) for the new starred aggregates, and $\overline{\overline{x}}_1^* = \overline{\overline{x}}_1$, $\overline{\overline{x}}_2^* = \overline{\overline{x}}_2$, $\overline{\overline{x}}_3^* = \overline{\overline{x}}_3$, $\overline{\overline{x}}_4^* = \overline{\overline{x}}_4$. Thus if the theorem be proved for the starred aggregates, it holds for the original ones.

We have to shew that, after suitable transformations, a system of division of the aggregates into parts, of the form in (6), can be found, such that $\overline{\overline{x}}_{13} + \overline{\overline{x}}_{14} = \overline{\overline{x}}_{14}$, and $\overline{\overline{x}}_{31} + \overline{\overline{x}}_{32} = \overline{\overline{x}}_{32}$. For, from these equations, we deduce $\overline{\overline{x}}_1 = \overline{\overline{x}}_2 = \overline{\overline{x}}_{14}$, $\overline{\overline{x}}_3 = \overline{\overline{x}}_4 = \overline{\overline{x}}_{23}$, and then the aggregates x_2, x_4 are such that each has a part which is equivalent to the other; and consequently, in accordance with the equivalence theorem, x_2, x_4 are equivalent to one another, or $\overline{\overline{x}}_2 = \overline{\overline{x}}_4$. It has in fact to be shewn that x_{13} can be so chosen that it is negligible with respect to cardinal number, in comparison both with x_{14} and with x_{23}.

We form the systems of transformations

$$\phi_b = \chi_2,\ \phi_b\phi_c = \chi_4,\ \phi_b\phi_c\phi_b = \chi_6,\ \dots,$$
$$\phi_c = \chi_3,\ \phi_c\phi_b = \chi_5,\ \phi_c\phi_b\phi_c = \chi_7,\ \dots;$$

each transformation χ in this system has one inverse, given by the scheme $\chi_{4n}\chi_{4n+1} = 1$, $\chi_{4n+2}\chi_{4n+2} = 1$, $\chi_{4n+3}\chi_{4n+3} = 1$; thus the transformations χ form a group of reversible transformations of M ($\equiv (x_1, x_2)$) into itself.

An element e_{13}, of x_{13}, is either, (i) transformed into an element of x_{24} by a transformation χ, with finite index, or else, (ii) e_{13} is not transformed into an element of x_{24} by any of the transformations χ. Suppose then that, for every element e_{13}, of x_{13}, the second of these cases arises, then χ_{2r}, χ_{2r-1} transform the elements of x_{13} into aggregates which are respectively in x_{23} and x_{14}, and in them these aggregates form an enumerable system of separate parts of each. For, in the case contemplated, χ_2 transforms x_{13} into a part of x_{23}; by χ_4, the elements of x_{13} become elements of x_{23} or x_{24}, consequently, in accordance with (ii), $\chi_4(x_{13})$ is a part of x_{23}. In this manner it is seen that x_{13} is transformed, by every χ_{2n}, into a part of x_{23}, and by every χ_{2n+1}, into a part of x_{14}. It then follows, by proposition (5), that $\overline{\overline{x}}_{13} + \overline{\overline{x}}_{14} = \overline{\overline{x}}_{14}$, $\overline{\overline{x}}_{13} + \overline{\overline{x}}_{23} = \overline{\overline{x}}_{23}$, and the theorem is thus completely established. The remainder of the proof consists in shewing that, by an exchange of elements of x_{13} with elements of x_{24}, it is possible to arrange so that the case just considered always arises.

Suppose x_{13}' are those elements of x_{13} which are transformed by χ_2 into elements of x_{24}; let x_{13}'' denote those elements, different from x_{13}', which are transformed, by χ_3, into elements of x_{24} which were not affected by χ_2, and so on; we thus obtain the scheme

$$\begin{cases} x_{13}' & \chi_2(x_{13}') & \text{in } x_{24}, \\ x_{13}' \neq x_{13}'' & \chi_2(x_{13}') \neq \chi_3(x_{13}'') & \text{in } x_{24}, \\ x_{13}' \neq x_{13}'' \neq x_{13}''' & \chi_2(x_{13}') \neq \chi_3(x_{13}'') \neq \chi_4(x_{13}''') & \text{in } x_{24}, \\ x_{13}' \neq \dots \neq x_{13}^{(n)} & \chi_2(x_{13}') \neq \chi_3(x_{13}'') \neq \dots \neq \chi_{n+1}(x_{13}^{(n)}) & \text{in } x_{24}. \end{cases}$$

We take now the equivalent sums

$$[x_{13}] = \sum_{n=1}^{\infty} x_{13}^{(n)}, \text{ and } [x_{24}] = \sum_{n=1}^{\infty} \chi_{n+1}(x_{13}^{(n)}),$$

and we carry out an exchange of $[x_{13}]$ with $[x_{24}]$; we then have

$$x_{13} = [x_{13}] + [(x_{13})], \quad x_{24} = [x_{24}] + [(x_{24})].$$

When the exchange has been made, of the elements of $[x_{13}]$ with those of $[x_{24}]$, we denote the new aggregates by starring the original ones; we have then, in accordance with the formulae (6), expressions for x_1^*, x_2^*, x_3^*. x_4^*, and we can, as has been shewn above, attend to these, instead of to the original x_1, x_2, x_3, x_4. Now no element of x_{13}^* is transformed into an element of x_{24}^* by any of the transformations χ, it being understood that the transformations χ are not to affect the substituted elements; and thus, by the reasoning which has been given above, for the case in which no element of x_{13} is transformed into an element of x_{24}, the theorem is established. Bernstein has also proved that, if $2\alpha = \alpha + \beta$, where α, β are cardinal numbers, then $\alpha \geqq \beta$.

THE ORDER-TYPE OF SIMPLY ORDERED AGGREGATES

157. *An aggregate M is said to be a simply ordered aggregate when each element m has a definite rank relatively to the other elements of M, so that, of any two elements m, m' whatever, it is known which has the higher and which has the lower rank.*

If m has a lower rank than m', the fact is denoted symbolically by $m < m'$; and if a higher rank, by $m > m'$.

If an aggregate is given, at first unordered, it may be possible to order the aggregate in a variety of essentially distinct ways. If the aggregate is finite, the ordering of it may be accomplished by arbitrarily assigning to each element its rank relatively to the others. In case the aggregate is an infinite one, the ordering of it consists in the setting up of some general rule which suffices logically to assign the relative order of any two elements.

Besides simply ordered aggregates there exist also doubly, or trebly, ordered aggregates, or also aggregates with higher degrees of multiplicity[†] of order. Each element of such an aggregate possesses two, three, or more distinct characteristics of an ordinal character. Only simply ordered aggregates will be considered here.

Two simply ordered aggregates M, N are said to be similar when a (1, 1) *correspondence can be established, in accordance with some law, such that, to any two definite elements m, m', of M, there correspond two definite elements n, n', of N, in such a manner that the relative order of m, m', in M, is the same as that of the corresponding elements n, n', in N.*

[†] Multiple order-types have been considered by F. Riesz, *Math. Annalen*, vol. LXI (1905), p 406.

This relation of similarity may be represented symbolically by $M \simeq N$.

Every simply ordered aggregate is similar to itself.

Two simply ordered aggregates which are similar to a third are similar to one another.

All simply ordered aggregates which are similar to one another are said to have the same order-type.

An order-type is accordingly characteristic of a class of similar aggregates.

The order-type of a simply ordered aggregate M is defined by Cantor as the concept which is obtained by abstraction when the nature of the elements of M is disregarded, their order being alone retained. The order-type of M is then denoted by $\bar{M}$. This definition will be further discussed in § 191. That similar aggregates have the same order-type is regarded by Cantor as a deduction from this definition.

If, in $\bar{M}$, we further disregard the order of the elements, we obtain $\bar{\bar{M}}$, the cardinal number of M.

The order-type of M is, from Cantor's point of view, regarded as a simply ordered aggregate similar to M, such that each element is the number 1. If any order-type be denoted by α, the corresponding cardinal number is denoted by $\bar{\alpha}$.

Corresponding to any given transfinite cardinal number, there is a multiplicity of order-types which form a class; each such class of order-types is characterized by the common cardinal number of all the order-types of the class.

The order-types which belong to the class corresponding to a cardinal number a form an aggregate which has a cardinal number a'. It will appear that a' is always greater than a.

If the order of every pair of elements in a simply ordered aggregate M be reversed, the aggregate in the new order is denoted by $^*M$.

If $\bar{M} \equiv \alpha$, then the order-type $^*\bar{M}$ is denoted by $^*\alpha$.

The order-type of the aggregate of all the finite integers, in their natural order $(1, 2, 3, \ldots)$, is denoted by ω. This is therefore the order-type of every aggregate $(a_1, a_2, \ldots a_n, \ldots)$ which is similar to $(1, 2, 3, \ldots)$.

The aggregate $(\ldots a_n \ldots a_3, a_2, a_1)$ has the order-type $^*\omega$.

THE ADDITION AND MULTIPLICATION OF ORDER-TYPES

158. If M, N denote two simply ordered aggregates, and if the aggregate (M, N) be formed, in which all the elements of both M and N occur, and which is such that any two elements of M have the same relative order as in M, and that any two elements of N have the same relative order as in N, and further that each element of M has a lower rank than all the elements of N, then the new simply ordered aggregate (M, N) is said to be the *sum* of the two simply ordered aggregates M and N. It is clear that if $M \simeq M'$, $N \simeq N'$, then $(M, N) \simeq (M', N')$, and thus that the order-type of (M, N) depends only on the order-types of M and N.

If $\overline{M} = \alpha$, $\overline{N} = \beta$, *the sum* $\alpha + \beta$ *is defined to be the order-type of the sum* (M, N) *of the two simply ordered aggregates, as defined above.*

This defines the operation of addition of order-types. It will be seen that the addition of order-types does not obey the commutative law. For if $\alpha = \overline{M}$, $\beta = \overline{N}$, then $\alpha + \beta = (\overline{M, N})$: but $\beta + \alpha = (\overline{N, M})$; and the two order-types $(\overline{M, N})$, $(\overline{N, M})$ are in general different from one another.

If n denotes a finite integer, $\omega + n$ is the order-type of the ordered aggregate $(e_1, e_2, e_3, \ldots f_1, f_2, \ldots f_n)$, whereas $n + \omega$ is the order-type of $(f_1, f_2, \ldots f_n, e_1, e_2, e_3. \ldots)$. It is clear that the first of these aggregates is not similar to $(g_1, g_2, g_3, \ldots)$, but if we make $f_1, f_2, \ldots f_n$ correspond to $g_1, g_2, g_3, \ldots g_n$, then e_1 to g_{n+1}, e_2 to g_{n+2}, ... and in general e_m to g_{n+m}, it is seen that the second of the above order-types is similar to $(g_1, g_2, g_3, \ldots)$. It thus appears that $n + \omega = \omega$, but $\omega + n \neq \omega$.

159. In the simply ordered aggregate N, let us suppose that, in the place of each element, there is substituted a simply ordered aggregate similar to M, whereby a new simply ordered aggregate is formed; this may be denoted by $M.N$. It is clear that if $M \simeq M'$, $N \simeq N'$, then $M.N \simeq M'.N'$; thus the order-type of $M.N$ depends only on the order-types of M and N.

If $\alpha = \overline{M}$, $\beta = \overline{N}$, *the product* $\alpha.\beta$ *is defined to be* $\overline{M.N}$, *the order-type of* $M.N$, *as just defined.*

It will be seen that the product $\alpha.\beta$ is in general different from $\beta.\alpha$, and thus that the multiplication of order-types does not obey the commutative law. For example, $\omega.2$ is the order-type of the aggregate formed by substituting in (a_1, a_2) for each of the two elements an aggregate of type ω; $\omega.2$ is therefore the order-type of $(b_1, b_2, b_3, \ldots c_1, c_2, c_3, \ldots)$, in which there is no last element, and no element immediately preceding c_1. On the other hand, $2.\omega$ is the order-type obtained by substituting for each element in $(a_1, a_2, a_3, \ldots)$, an aggregate consisting of two elements; and $2.\omega$ is thus the order-type of the enumerable aggregate

$$(a_{11}, a_{12}, a_{21}, a_{22}, a_{31}, a_{32}, \ldots),$$

which is similar to $(b_1, b_2, b_3, \ldots)$, as may be seen by making a_{n1} correspond to b_{2n-1} and a_{n2} to b_{2n}. It has thus been shewn that $2.\omega = \omega$, but $\omega.2 \neq \omega$.

THE STRUCTURE OF SIMPLY ORDERED AGGREGATES

160. An examination of the structure of a simply ordered aggregate M can, in general, only be attempted by considering the nature of those aggregates which are its parts, and in each of which parts the order of the elements is the same as that of the same elements in the whole aggregate. The simplest transfinite part of an ordered aggregate is that which has one of the types ω, $*\omega$. Such parts we speak of as *ascending sequences*, and *descending sequences*, respectively, contained in M.

Two ascending sequences $\{a_n\}$, $\{a_n'\}$, contained in M, are said to be *related to one another*, provided that, corresponding to any element a_n, of the first, there are elements $a'_{n'}$, of the second, such that $a_n < a'_{n'}$; and provided also that, corresponding to any element a_n', of the second, there are elements $a_{n'}$, of the first sequence, such that $a_n' < a_{n'}$.

Two descending sequences $\{b_n\}$, $\{b_n'\}$, contained in M, are said to be related to one another, provided that, corresponding to any element b_n of the first sequence, there are elements b'_n, of the second, such that $b_n > b'_{n'}$, and provided also that, corresponding to any element b_n', of the second, there are elements $b_{n'}$, of the first sequence, such that $b_n' > b_{n'}$.

An ascending sequence $\{a_n\}$ and a descending sequence $\{b_n\}$, contained in M, are said to be related to one another, if $a_n < b_{n'}$ for every n and n': and further, provided there exists in M no element, or only one element m, which is such that $a_n < m < b_n$, for every n.

Two sequences contained in an ordered aggregate, which are both related to a third sequence, are related to one another.

Two sequences in an ordered aggregate, which are both ascending, or both descending, and of which one is a part of the other, are related to one another.

161. Suppose that, in an ordered aggregate M, there is an element m_0 which satisfies the following conditions, with respect to an ascending sequence contained in M:

(1), for every n, $a_n < m_0$;

(2), for every element m, of M, which is $< m_0$, there exists a number n such that $a_n, a_{n+1}, a_{n+2}, \ldots$ are all $> m$; then the element m_0 is said to be the *limiting element*, or *limit of* $\{a_n\}$ *in* M; and m_0 is said to be a *principal element* of M.

Similarly, if we suppose that, in M, there is an element m_0, which satisfies with reference to a descending sequence $\{a_n\}$, contained in M, the following conditions:

(1), for every n, $a_n > m_0$;

(2), for every element m, of M, which is $> m_0$, there exists a number n

such that $a_n, a_{n+1}, a_{n+2}, \dots$ are all $< m$; then the element m_0 is said to be the *limiting element*, or *limit of* $\{a_n\}$ *in* M; and m_0 is said to be a *principal element* of M.

A sequence contained in M can never have more than one limiting element in M.

If a sequence in M has a limiting element m_0, in M, then m_0 is the limiting element of every sequence in M which is related to the first one.

Two sequences which have the same limiting element, in M, must be related to one another.

It is clear that, if M, M' are similar ordered aggregates, an ascending, or a descending, sequence in M corresponds to a sequence of the same kind in M'. To every principal element in M there corresponds a principal element in M'.

An ordered aggregate which is such that every element is a principal element is said to be dense in itself.

If, in an ordered aggregate, every sequence which is contained therein has a limiting element in the aggregate, then the ordered aggregate is said to be a closed aggregate.

An ordered aggregate which is dense in itself, and also closed, is said to be perfect.

An ordered aggregate which is such that, between any two whatever of its elements, there are other elements of the aggregate, is said to be everywhere dense.

The properties of an ordered aggregate, thus defined, are also properties of any similar aggregate; hence the terms may be applied to the order-types which are symbolized by replacing the elements of the ordered aggregates by 1; there can exist therefore an order-type which is dense in itself, or closed, or perfect, or everywhere dense.

The terms which have been here employed for the purpose of describing certain peculiarities which may exist in an ordered aggregate, or in the corresponding order-type, are identical with those which we have employed in analogous senses in Chapter II, in the case of sets of points, or numbers. There is however a distinction which must be noticed between the use of the terms in the two cases. To illustrate this distinction, let $(P_1, P_2, P_3, \dots P_n, \dots)$ be a sequence of points on a straight line, which sequence has a limiting point P_ω, on the right of the points P_n; then if Q be any point of the straight line on the right of P_ω, the two ordered aggregates $(P_1, P_2, P_3, \dots P_n, \dots P_\omega)$, and $(P_1, P_2, P_3, \dots P_n, \dots Q)$, are similar, and have the same order-type $\omega + 1$. In the first of these aggregates, P_ω is the limiting element of the sequence $(P_1, P_2, \dots P_n, \dots)$; and, in the second aggregate, Q is the limiting element of the same sequence; and therefore both the ordered aggregates are closed, in the sense explained above. The first of these aggregates forms a closed set of points, in the sense of the term defined in Chapter II; but the second does not, since Q is not a limiting point of the set of points $\{P_n\}$ The distinction rests upon the different use of the terms limiting element and limiting point, in the two cases of an ordered aggregate of elements in general, and that of a set of points in the continuum. The question whether an element is a limiting element of an aggregate to which it belongs, or not, in the sense defined above, is answered by examining the structure of the ordered aggregate itself. In the case of a set of points in the continuum, a particular point may be a limiting element of the aggregate of points considered merely as an aggregate of elements with a particular order-type; but the question as to whether the same point is a limiting point of the set of points, considered as chosen out of the continuum, can only be answered after an examination of the ordinal relation of the point to other points of the continuum which do not belong to the set, in fact, the set must be regarded, for this purpose, as an aggregate which is only a part of another aggregate, the continuum. It is now clear that a set of points, considered solely as an ordered aggregate of elements, without reference to the fact that it is essentially a part of the continuum, may be closed, or perfect; and yet that the same set of points need be neither closed nor perfect, in the sense of the terms employed in the theory of sets of points, which has been dealt with in Chapter II.

THE ORDER-TYPES η, θ, π

162. Certain order-types which are of special importance will be now examined.

The first of these is the order-type η, of the set R, of rational numbers between 0 and 1 (both exclusive), in their order, as defined in Chapter I.

It will be shewn that the order-type η is exhaustively characterized by the following properties:

(1). $\bar{\bar{\eta}} = \aleph_0 = a$.

(2). There is in η no lowest, and no highest, element.

(3). η is everywhere dense.

In fact, every simply ordered aggregate M, which has these three characteristics, is similar to the aggregate R.

To prove this, we first observe that, on account of the condition (1), the order of the elements in both M and R can be so altered that each of them is reduced to the order-type ω. Let this be done; and denote by M_0, R_0 the new ordered aggregates

$$M_0 = (m_1, m_2, m_3, \ldots),$$
$$R_0 = (r_1, r_2, r_3, \ldots).$$

We have to shew that $M \simeq R$; and to do this we have to shew how to establish the requisite correspondence between the elements m, of M, and r, of R. Let m_1 be made to correspond to r_1; then there are an indefinitely great number of elements of M, which have the same relation, as regards order, to m_1, as r_2 has, in R, relatively to r_1; of all these elements choose that one m_{ϵ_2}, which has the smallest index as it appears in M_0; and let m_{ϵ_2} be made to correspond to r_2. Of all the elements of M, which are related to m_1 and m_{ϵ_2}, in the same manner, as regards order in M, as r_3 is related to r_1 and r_2, as regards order in R, choose that one m_{ϵ_3}, which has the smallest index as it appears in M_0; and make m_{ϵ_3} correspond to r_3.

Proceeding in this manner, we make the elements $r_1, r_2, r_3, \dots r_n$, of R, correspond to the elements $m_1, m_{\epsilon_2}, m_{\epsilon_3}, \dots m_{\epsilon_n}$, of M; and so far as these elements are concerned the relations of rank are preserved in the correspondence: we proceed then to choose, in the same manner as before, the element $m_{\epsilon_{n+1}}$, which is to be made to correspond to r_{n+1}, and thus we obtain, for every r_n, the corresponding m_{ϵ_n}. It must however be shewn that this process exhausts all the elements m, of M, that is to say, that in the sequence $1, \epsilon_2, \epsilon_3, \dots \epsilon_n, \dots$ every integral number p occurs in some definite place. This can be proved by the method of induction. Let us assume that the elements $m_1, m_2, m_3, \dots m_n$ all occur in the correspondence that has been set up between the whole of R and at least a part of M; then we shall prove that m_{n+1} also occurs. Upon this assumption, let λ be so great that, among the elements $m_1, m_{\epsilon_1}, m_{\epsilon_2}, \dots m_{\epsilon_\lambda}$, all the elements $m_1, m_2, m_3, \dots m_n$ occur. Then, if m_{n+1} is not also among those elements, choose out of $r_{\lambda+1}, r_{\lambda+2}, r_{\lambda+3}, \dots$ that element $r_{\lambda+s}$, with the smallest index, which has the same relation to $r_1, r_2, \dots r_\lambda$, as regards order in R, that m_{n+1} has relatively to $m_1, m_{\epsilon_1}, m_{\epsilon_2}, \dots m_{\epsilon_\lambda}$, as regards order in M. Then the element m_{n+1} has the same relation to $m_1, m_{\epsilon_1}, m_{\epsilon_2}, \dots m_{\epsilon_{\lambda+s-1}}$, as regards order in M, as $r_{\lambda+s}$ has to $r_1, r_2, \dots r_{\lambda+s-1}$, as regards order in R. It thus appears that m_{n+1} is the element with the smallest index as it appears in M_0, which has, in M, the same relation as regards order to $m_1, m_{\epsilon_1}, \dots m_{\epsilon_{\lambda+s-1}}$, that $r_{\lambda+s}$ has, relatively to $r_1, r_2, \dots r_{\lambda+s-1}$, in R; hence $m_{\epsilon_{\lambda+s}} \equiv m_{n+1}$; that is, the element m_{n+1} occurs in the correspondence which has been established between M and R. It has now been shewn that M and R are similarly ordered aggregates.

Examples of the order-type η are the following:

(1). The aggregate of all negative and positive rational numbers, including zero, in their natural order.

(2). The aggregate of all rational numbers which are greater than a, and less than b, where a, b are two real numbers such that $a < b$.

(3). The aggregate of all real algebraical numbers in their natural order in the continuum, or of all such of these numbers as lie between two real numbers a, b.

(4). The aggregate of a set of non-abutting linear intervals which are such that their end-points and the limiting points of these end-points form a non-dense perfect set of points in a linear interval.

The rational numbers of the interval (0, 1), including 0 and 1, form an aggregate of the order-type $1 + \eta + 1$.

163. We now proceed to the consideration of the order-type θ, of points forming a linear closed continuum.

It will be shewn that any simply ordered aggregate M is similar to the aggregate X of all real numbers of the continuum (0, 1), in their natural order, provided (1), M is perfect, and (2), in M, an aggregate S, with the cardinal number $\aleph_0$, is contained, which is so related to M, that between any two elements m_0, m_1, of M, there are elements of S.

If S has a lowest and a highest element, these can be removed without affecting its relation to M; and thus we may suppose S to be of the type η. of the aggregate R of rational numbers which lie between 0 and 1, both exclusive, in their natural order.

Since $S \simeq R$, we may suppose the elements of S to be made to correspond in order to the elements of R; and it will be shewn that this correspondence enables us to establish a correspondence between the elements of M and X.

We suppose that each element of M, which belongs to S, corresponds to that element of X which belongs to R, just as in the correspondence of S with R already established. Any element m, of M, which does not belong to S, is the limiting element of a sequence $\{m_n\}$ of elements of S To this sequence $\{m_n\}$, there corresponds a sequence $\{r_n\}$ in X, all the elements of which belong to R, and this sequence $\{r_n\}$ has a limiting element x, in X, not belonging to R; we take therefore m, in M, to correspond to x, in X. If we take a different sequence $\{x_n'\}$, which has the same limiting element m as before, in M, then there corresponds to it a sequence $\{r_n'\}$ in R, which has the same limiting element x as before, in X. It will now be shewn that, in the correspondence so established between the elements of M and of X, the relative order of two elements of M is the same as that of the corresponding elements of X. This clearly holds of any two elements of M which are also elements of S. Consider next two elements m and s, of M, the first of which does not, and the second of which does, belong to S; and let x_1, r be the corresponding elements of X. If $r < x_1$, there exists an ascending sequence in R, of which x_1 is the limiting element, such that all its elements are $> r$; then, to this sequence there corresponds an ascending sequence in S, all the elements of which are $> s$, and of which m is the limiting element; hence $s < m$. If $r > x_1$, it can be shewn, in a similar manner, that $s > m$. The proof that, corresponding to any two elements m_1, m_2, of M, which do not belong to S, the elements x_1, x_2, of X, are such that $m_1 \gtrless m_2$, according as $x_1 \gtrless x_2$, is of a precisely similar character to that just given. It has thus been shewn that M and X are similar aggregates, and that the type θ is characterized by the conditions (1) and (2).

The above characterization[†] of the type θ contains Cantor's ordinal theory of the constitution of the linear closed continuum.

A non-dense perfect set of points in a linear interval has not the order-type θ, but the set of complementary intervals together with the limiting points of their end-points does form an aggregate of order-type θ, when the elements, consisting partly of points, and partly of intervals, are taken in the order in which they occur in the continuum.

164. The order-type $^*\omega + \omega$ may be denoted by π, and is the order-type of the negative and positive integers in their natural order. This order-type has properties distinct from those of ω. For example, $n + \omega$ has been shewn to be identical with ω, where n is a finite integer; but $n + \pi$ is not identical with π. From either of the equations $n + \pi = m + \pi$, or $\pi + n = \pi + m$, there follows $m = n$, or more generally[‡]:

If n, n' are finite integers, ζ and ζ' other order-types, from the equation $n + \pi + \zeta = n' + \pi + \zeta'$, there follows $n = n'$, $\zeta = \zeta'$.

To prove this theorem, we observe that, if the two aggregates be placed into similar correspondence, the lowest elements correspond to one another, then the second, and so on; hence $n = n'$ is proved at once: and we now have $\pi + \zeta = \pi + \zeta'$.

When two simply ordered aggregates $M_\pi + Z$, $N_\pi + Z'$, of order-types $\pi + \zeta$, $\pi + \zeta'$, are placed in correspondence in order, either M_π corresponds to N_π, or M_π corresponds to a part of N_π, or else N_π corresponds to a part of M_π. In the last two cases the order-type π must be split up into $\pi = \pi_1 + \pi_2$, where $\pi = \pi_1$, and π_2 is some other order-type; but from the definition $\pi = {}^*\omega + \omega$, it is clear that every mode of dividing π into two parts, without altering the relative order of the elements, leaves it in the form $^*\omega + \omega$; hence it is impossible that $\pi = \pi_1 + \pi_2$, and $\pi = \pi_1$; and therefore M_π corresponds to N_π. Hence also Z corresponds to Z'; or $\zeta = \zeta'$.

[†] See Russell, *Principles of Mathematics*, vol. I (1903), p. 303, also Veblen, *Trans. Amer. Math. Soc.* vol. VI (1905), p. 165 and Huntington, *Annals of Math.* (2), vols. VI (1904), p. 151 and VII (1905), p. 15.

[‡] Bernstein, *loc. cit.* p. 9.

NORMALLY ORDERED AGGREGATES

165. The order-type of a simply ordered aggregate is, as we have already seen, such that the structure of the aggregate, as revealed by an examination of the sequences contained in it, may be of the most varied character; the various sequences may be ascending or descending ones, and may, or may not, have a limiting element within the aggregate.

Of all the possible order-types, those are of especial importance which have been defined by Cantor as the order-types of normally ordered aggregates (wohlgeordnete Mengen).

A normally ordered aggregate M is one which satisfies the following conditions:

(1). *M has an element m, of lower rank than all the other elements.*

(2). *If M_1 is any part of M, and if M contains one or more elements which are of higher rank than all the elements of M_1, then there exists one element m', of M, which immediately follows the part-aggregate M_1, so that there are no elements of M which are intermediate in rank between m' and all the elements of M_1.*

The special case of (2) which arises when M_1 consists of one element, shews that a normally ordered aggregate is such that each element has one which immediately follows it, unless the element is the highest element of M. It is however not necessarily the case that M has a highest element.

If $e_1, e_2, e_3, \dots e_n, \dots$ be an ascending sequence of elements contained in M, and such that elements exist, in M, which are of higher rank than every e_n, then there exists an element e', of M, which is higher than all the e_n, and such that every element e'', of M, which is lower than e' is lower than $e_n, e_{n+1}, e_{n+2}, \dots$, for some definite value of n.

Every part of a normally ordered aggregate has a lowest element.

Let M_1 be a part of M; if M_1 contains m_1, the lowest element of M, then m_1 is the lowest element of M_1. If M_1 does not contain m_1, consider that part of M which contains all those elements every one of which is of lower rank than all the elements of M_1; this part of M must have an element which immediately follows it; and this element belongs to M_1, and is its lowest element.

If a simply ordered aggregate M itself, and also every part of M, has a lowest element, then M is normally ordered.

The condition (1) is fulfilled. Let M_1 be a part of M such that M contains elements which are higher than all those of M_1; let these form the aggregate M_2, and let m be the lowest element of M_2. Then m is the element which immediately follows M_1; and thus the condition (2) is satisfied.

This property of a normally ordered aggregate, that every part of it as a lowest element, might be adopted as the definition of a normally ordered aggregate.

A somewhat simpler property, which might be employed to define a normally ordered aggregate, is the following:

An aggregate M is normally ordered[†] *if, and only if, it contains no part, of which the order-type is* $^*\omega$.

If M is not normally ordered, at least one part of it must have no lowest element, and this part contains a sequence whose order-type is $^*\omega$. This follows from the theorem of § 151, that every aggregate that is not finite contains a part, of cardinal number $\aleph_0$. Such part can be ordered according to the type ω, when it has no highest element, and according to the type $^*\omega$, when it has no lowest element. An aggregate which has a lowest element, and is also such that each element has one that immediately succeeds it, is not necessarily normally ordered, even if each element has one immediately preceding it. This can be seen by considering an aggregate with the order-type $\omega + {}^*\omega$.

166. The following properties of normally ordered aggregates can be proved in a very simple manner:

Every part-aggregate of a normally ordered aggregate is itself normally ordered.

Every ordered aggregate which is similar to a normally ordered aggregate is itself normally ordered.

If, in a normally ordered aggregate M, there be substituted for the elements normally ordered aggregates, in such a manner that, if M_m, $M_{m'}$ are the aggregates substituted for any two elements m, m', then $M_m \lesseqgtr M_{m'}$, according as $m \lesseqgtr m'$, the resulting new aggregate is normally ordered.

167. *The part of a normally ordered aggregate M which consists of all those elements which are of lower rank than an element m, of M, is called the segment of M determined by the element m.*

The aggregate which remains when the segment of M, determined by the element m, is removed from M, is called the *remainder* of M determined by an element m. The element m is the lowest element of the remainder.

If S is the segment of M, determined by m, and R is the remainder, then $M = (S, R)$.

The segment of M determined by the lowest element of M contains no element, but it may be regarded as existent, and being the null-segment. The remainder, determined by the lowest element, is the aggregate M itself.

[†] See Jourdain, *Phil. Mag.* (6), vol. VII (1904), p. 65.

Of two segments S, S' determined by the elements m, m', of which $m < m'$, we say that S is the smaller, and S' the larger segment, or $S < S'$.

It can easily be seen that, if M, M_1 are two similar normally ordered aggregates, a segment of M corresponds to a similar segment of M_1, the element by which the segment of M is determined corresponding to the element of M_1 by which the segment of M_1 is determined.

A normally ordered aggregate is not similar to any of its segments.

Assume that, if possible, $S \simeq M$, and suppose the elements of S, M are put into correspondence. To the segment S, of M, there must correspond a segment S_1, of S, so that $S_1 \simeq M \simeq S$, where $S_1 < S$. Since $S_1 \simeq M$, we find in a similar manner a segment $S_2 < S_1$, which is similar to M, and so on; and in this way we obtain an unending sequence

$$S > S_1 > S_2 > \ldots > S_n > \ldots$$

of segments of M, which are all similar to M. Let $m, m_1, m_2, \ldots m_n, \ldots$ be the elements which determine the segments S, S_1, S_2, ... S_n, ...; then $m > m_1 > m_2 > \ldots > m_n > \ldots$.

The aggregate $(\ldots\ldots m_n, \ldots m_2, m_1, m)$ would be a part of M which has no lowest element, and is of type $^*\omega$; this is impossible if M is normally ordered.

If M is an infinite normally ordered aggregate, it always has parts which are similar to M, although such a part cannot be a segment.

A normally ordered aggregate cannot be similar to any part of one of its segments.

Let us assume that, if possible, S' a part of a segment S, of M, is similar to M. Since $S' \simeq M$, we can place the elements of S', M in correspondence, then, to the segment S, of M, there will correspond a segment S_1, of S', where $S_1 \simeq S$; let then S_1 be determined by the element e_1, of S'. Since e_1 is also an element of M, it determines a segment M_1, of M, of which S_1 is a part, and which has a part similar to M. Proceeding in the same manner, we determine a segment M_2, of M_1, which has a part that is similar to M; and in this way we obtain an unending sequence of segments of M, all similar to M, so that $M > M_1 > M_2 > \ldots > M_n > \ldots$. The elements which determine these sequences form a part of M which is of type $^*\omega$; and this is contrary to the hypothesis that M is normally ordered.

Two different segments of a normally ordered aggregate cannot be similar.

For one of these segments is a segment of the other.

The aggregate of which the elements are the segments of a normally ordered aggregate M, can be so ordered that it is similar to M.

For each element of M corresponds to a single segment of M, and thus the aggregate of segments can be so ordered as to be similar to M.

There is only one mode of putting the elements of two similar normally rdered aggregates into correspondence, so that the relative orders of the lements are unaltered in the correspondence.

For if, in two modes of placing the aggregates in correspondence, two lements f, f', of one aggregate M, correspond to one element e of the ther M', the segments of M determined by f, f' are each similar to the egment of M' determined by e; but it has been shewn to be impossible hat M can have two different segments which are similar to one another.

A segment of one of two normally ordered aggregates is similar to at most me segment of the other aggregate.

If S, S' are similar segments of two normally ordered aggregates M, M', then, to every smaller segment $S_1 < S$, of M, there corresponds a similar segment $S_1' < S'$, of M'.

If S_1, S_2 are two segments of the normally ordered aggregate M, and S_1', S_2' are two similar segments of a normally ordered aggregate M', then if $S_1' < S_1$, it follows that $S_2' < S_2$.

If a segment S, of M, is not similar to any segment of another normally ordered aggregate M', then no segment $S' > S$, of M, is similar to any segment of M', nor to M' itself; and the same holds of M itself

If M, M', two normally ordered aggregates, are so related that, to any segment of either, there corresponds a similar segment of the other, then $M \simeq M'$.

Any element e, of M, determines a segment of M which corresponds to a similar segment of M'. Let this latter be determined by an element e', of M'; we then take e to correspond to e'. To every element of M we therefore find a corresponding element of M', and it is seen by applying the foregoing theorems that the relative order of the elements is preserved.

168. *If two normally ordered aggregates M, M' are so related that,* (1), *to every segment S, of M, there corresponds a similar segment S', of M', and* (2), *at least one segment of M' exists, to which there is no corresponding similar segment of M; then there exists a definite segment S_1', of M', such that $S_1' \simeq M$.*

Consider all those segments of M' which do not correspond to similar segments of M. Among these, there must be one S_1', which is the least of all; this follows from the fact that the elements which determine these segments of M' form an aggregate which has a lowest element, and this lowest element determines the segment S_1'. Every segment of M' which is greater than S_1' is such that there exists no corresponding similar segment of M; but every segment of M' which is less than S_1' has a corresponding similar segment of M. Since, to every segment of M, there corresponds a similar segment of S_1', and, to every segment of S_1', there corresponds a similar segment of M, it follows that $M \simeq S_1'$.

If the normally ordered aggregate M' has at least one segment to which there corresponds no similar segment of M, then, to every segment of M, there corresponds a similar segment of M'.

Let S_1' be the smallest segment of M' to which there corresponds no similar segment of M. If there exist segments of M to which no corresponding similar segments of M' exist, let S_1 be the smallest of all such segments of M. To every segment of S_1 there corresponds a similar segment of S_1', and conversely; hence $S_1 \simeq S_1'$, which is contrary to the hypothesis that there exists no segment of M which is similar to S_1'.

If M, M' are any two normally ordered aggregates, then either (1), M and M' are similar, or, (2), there exists a segment S', of M', which is similar to M, or, (3), there exists a segment S, of M, which is similar to M'; and these possibilities are mutually exclusive.

The following four possibilities may be contemplated, as regards the relation of M to M':

(1). To every segment of either M or M' there corresponds a similar segment of the other aggregate.

(2). To every segment of M there exists a corresponding similar segment of M'; but there is at least one segment of M' to which no similar segment of M corresponds.

(3). To every segment of M' there corresponds a similar segment of M; but there is at least one segment of M to which no similar segment of M' corresponds.

(4). There is at least one segment of M to which no similar segment of M' corresponds, and also at least one segment of M' to which no similar segment of M corresponds.

It has been shewn that (4) is impossible. In the case (1), it has been proved that $M \simeq M'$. In the case (2), it has been shewn that a definite segment S_1', of M', exists, such that $S_1' \simeq M$; and in the case (3), that there is a definite segment S_1, of M, such that $S_1 \simeq M'$.

It is impossible that, at the same time $M \simeq M'$, and also $M \simeq S_1'$; for, in that case, $M' \simeq S_1'$; and it has been shewn to be impossible that M' should be similar to one of its own segments.

It is also impossible that $M \simeq S_1'$, and also $M' \simeq S_1$; for there must then exist a segment of S_1' which is similar to S_1, and therefore to M'; but this is contrary to the theorem that a normally ordered aggregate cannot be similar to one of its segments.

If any part of M is such that that part is not similar to any segment of M, then that part is similar to M itself.

Any part M_1, of M, is normally ordered; if then M_1 be similar neither to M nor to any segment of M, there must exist a segment of M_1', of M_1, which is similar to M; and M_1' is a part of that segment of M which is determined by the same element that determines the segment M_1', of M_1. Therefore M would be similar to a part of one of its segments, which has been shewn to be impossible.

THE THEORY OF ORDINAL NUMBERS

169. *The order-type* $\bar{M}$, *of a normally ordered aggregate* M, *is said to be the ordinal number which belongs to* M; *all similar normally ordered aggregates have consequently the same ordinal number.*

If M, M' *are two normally ordered aggregates such that* M *has a segment which is similar to* M', *whilst* M' *has no segment which is similar to* M, *then the ordinal number* $\alpha \equiv \bar{M}$ *is said to be greater than the ordinal number* $\beta \equiv \bar{M}'$; *and this relation is denoted by* $\alpha > \beta$. *If* M *has no segment similar to* M', *but* M' *has a segment similar to* M, *the ordinal number* α *is said to be less than* β, *and the relation is denoted by* $\alpha < \beta$.

It follows from these definitions, in conjunction with the theorem of § 168, that, if α, β are any two ordinal numbers whatever, they satisfy one, and one only, of the relations $\alpha = \beta$, $\alpha > \beta$, $\alpha < \beta$; and that if $\alpha > \beta$, then $\beta < \alpha$.

Further it is seen that, if $\alpha < \beta$ and $\beta < \gamma$, then $\alpha < \gamma$; hence the aggregate of all ordinal numbers is a simply ordered aggregate, when arranged in such a manner that any one α, which has been defined as less than another one β, precedes it.

The sum $\alpha + \beta$ *of two ordinal numbers is, in accordance with the general definition of the sum of two order-types, the order-type of the normally ordered aggregate* (M, N), *where* M, N *are two normally ordered aggregates such that* $\alpha = \bar{M}$, $\beta = \bar{N}$.

Since M, N both contain no part of type $^*\omega$, the same is true of (M, N). Hence the aggregate (M, N) is normally ordered; and thus $\alpha + \beta$ is an ordinal number.

Since M is a segment of (M, N), we see that $\alpha < \alpha + \beta$.

N is a remainder of (M, N) determined by the lowest element of N, hence N may be similar to (M, N); or, if not, it is similar to a segment of (M, N): thus either $\beta = \alpha + \beta$, or $\beta < \alpha + \beta$.

The addition of ordinal numbers obeys the associative law, but not in general the commutative law; thus $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$, but $\alpha + \beta$ is in general not the same as $\beta + \alpha$.

170. The product $\alpha . \beta$, of two ordinal numbers, is, in accordance with the definition of § 159, the order-type of the aggregate obtained by substituting for each element of an aggregate of order-type β, an aggregate of order-type α. In accordance with the theorem of § 166, the aggregate thus obtained is normally ordered, and of type dependent only on α and β.

In general $\alpha . \beta$ is not equal to $\beta . \alpha$.

It is easily seen that $\alpha . \beta > \alpha$, provided $\beta > 1$; and that, if $\alpha\beta = \alpha\gamma$, then $\beta = \gamma$.

If α, β are two ordinal numbers such that $\alpha < \beta$, there exists an ordinal number γ such that $\alpha + \gamma = \beta$; and this number γ is defined to be $\beta - \alpha$.

For, if $\overline{M} = \beta$, there is a segment of M which may be denoted by M_1, such that $\overline{M}_1 = \alpha$; let then $M = (M_1, S)$, therefore $\overline{M} = \overline{M}_1 + \overline{S}$, and $\beta - \alpha = \overline{S}$.

171. Let $\beta_1, \beta_2, \ldots \beta_n, \ldots$ denote a simple sequence of ordinal numbers, and suppose $M_1, M_2, \ldots M_n, \ldots$ are aggregates of which the order-types are respectively the numbers of the sequence. The aggregate

$$(M_1, M_2, \ldots M_n, \ldots),$$

which is obtained by replacing each element of the normally ordered aggregate $(1, 1, 1, \ldots)$, of type ω, by a normally ordered aggregate, is, in accordance with the theorem of § 166, itself normally ordered; and its type defines the sum $\beta_1 + \beta_2 + \ldots + \beta_n + \ldots = \beta$. If α_n denotes the sum $\beta_1 + \beta_2 + \ldots + \beta_n$, we see that $\alpha_n = \overline{(M_1, M_2, \ldots M_n)}$; and it is clear that $\alpha_{n+1} > \alpha_n$: hence

$$\beta_1 = \alpha_1, \beta_2 = \alpha_2 - \alpha_1, \ldots, \beta_n = \alpha_n - \alpha_{n-1}.$$

It will now be shewn (1), that $\beta > \alpha_n$, for every value of n; and (2), that, if β' is any ordinal number $< \beta$, there is some definite value of n such that $\alpha_n, \alpha_{n+1}, \ldots$ are all $> \beta'$.

(1) follows from the fact that each α_n is the ordinal number of a segment of $(M_1, M_2, \ldots M_n, \ldots)$ of which β is the ordinal number.

To prove (2), we observe that a segment of $(M_1, M_2, \ldots M_n, \ldots)$ exists, of which β' is the ordinal number, and therefore the element which determines this segment must belong to one of the aggregates

$$M_1, M_2, \ldots M_n, \ldots,$$

say M_{n_1}. It follows that the segment is also a segment of $(M_1, M_2, \ldots M_{n_1})$; and therefore, $\beta' < \alpha_{n_1}$, or $\alpha_n > \beta'$, n being $\geq n_1$.

It has thus been proved that β is the ordinal number which immediately follows all the ordinal numbers $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$; and it may be spoken of as the limit of the sequence $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$.

Thus *every ascending sequence $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$ of ordinal numbers determines a limiting number $\beta = \lim\limits_{n \sim \infty} \alpha_n$, which immediately follows all the numbers of the sequence.*

THE ORDINAL NUMBERS OF THE SECOND CLASS

172. Every finite ordered aggregate is normally ordered, and its order-type is the ordinal number of the aggregate. The finite ordinal numbers may be spoken of as the ordinal numbers of the first class; to each such ordinal number there corresponds a single cardinal number, and the properties of the finite ordinal numbers are identical with those of the finite cardinal numbers, the terms ordinal and cardinal simply defining the two uses of the same number. In the case of transfinite aggregates there is no such identity between ordinal and cardinal numbers; in fact the arithmetic of the one kind of numbers is essentially different from that of the other kind.

Corresponding to a single transfinite cardinal number there is an infinity of transfinite ordinal numbers; all those transfinite ordinal numbers which correspond to aggregates that have one and the same cardinal number α are said to form a class $Z(\alpha)$, the class of normal order-types which have the cardinal number α.

The ordinal numbers of all those order-types which have the same cardinal number $\aleph_0$, as the aggregate of finite numbers, are said to be of the second class $Z(\aleph_0)$.

The ordinal number $\omega = \lim\limits_{n\sim\infty} n$, *and is the smallest number of the second class.*

If M denotes the aggregate $(m_1, m_2, \dots m_n, \dots)$, then $\overline{M} = \omega$, and $\overline{\overline{\omega}} = \aleph_0$. Any number β, which is $< \omega$, must be the order-type of a segment of M, and M has only segments $(m_1, m_2, \dots m_n)$ with finite ordinal numbers n; thus β must be a finite number; and therefore the only ordinal numbers $< \omega$ are finite ones.

Every number α, of the second class, has a number $\alpha + 1$ immediately following it.

For if $\alpha = \overline{M}$, $\overline{\overline{\alpha}} = \aleph_0$, we have $\alpha + 1 = (\overline{M, e})$, where e is a new element; and since M is a segment of (M, e), we have $\alpha + 1 > \alpha$. Also

$$\overline{\overline{\alpha + 1}} = \overline{\overline{\alpha}} + 1 = \aleph_0 + 1 = \aleph_0.$$

It has thus been shewn that $\alpha + 1$ is a number of the second class. Every number $< \alpha + 1$, is the order-type of a segment of (M, e); and such segment can only be M, or a segment of M; hence no number $< \alpha + 1$ is $> \alpha$: therefore $\alpha + 1$ is the next number greater than α.

If $\alpha_1, \alpha_2, \dots \alpha_n, \dots$ be any sequence of numbers of the second class, there is a number $\lim\limits_{n\sim\infty} \alpha_n$, also of the second class, which is the smallest number that is greater than every number α_n of the sequence.

If, as in § 171, we write $\beta_1 = \alpha_1,\ \beta_2 = \alpha_2 - \alpha_1, \ldots\ \beta_n = \alpha_n - \alpha_{n-1}, \ldots$, then if $\bar{G}_n = \beta_n$, we have $\lim\limits_{n\sim\infty} \alpha_n = \overline{(G_1, G_2, \ldots G_n, \ldots)}$; and this number $\lim\limits_{n\sim\infty} \alpha_n$ has been shewn to be the smallest number which is $> \alpha_n$ for every value of n. To shew that this number $\lim\limits_{n\sim\infty} \alpha_n$ is of the second class, we have, since $\overline{\beta}_n \leq \aleph_0$, for every value of n, $\lim\limits_{n\sim\infty} \alpha_n \leq \aleph_0 . \aleph_0 \leq \aleph_0$; and, since $\lim\limits_{n\sim\infty} \alpha_n$ is not finite, it must therefore $= \aleph_0$.

Two sequences $\{\alpha_n\}$, $\{\alpha'_n\}$, *of numbers of the second class, have the same limiting number, when, and only when, the sequences are related to one another, in accordance with the definition of* § 160.

Let β, γ be the two limiting numbers, and first assume that the sequences are related to one another. If $\beta < \gamma$, then for some value of n, $\alpha'_n > \beta$, $\alpha'_{n+1} > \beta, \ldots$; and hence, for some value of n', we must have $\alpha_{n'} > \beta$, $\alpha_{n'+1} > \beta, \ldots$, which is inconsistent with β being the limit of the sequence $\{\alpha_n\}$.

If we assume $\beta = \gamma$, then, since $\alpha_n < \gamma$, for some fixed number r we must have $\alpha'_r > \alpha_n$, $\alpha'_{r+1} > \alpha_n, \ldots$; and similarly, since $\alpha'_n < \beta$, for some fixed number s we must have $\alpha_s > \alpha'_n$, $\alpha_{s+1} > \alpha'_n, \ldots$; hence the two sequences are related to one another.

If n is a finite ordinal number, and α a number of the second class, then $n + \alpha = \alpha$, and hence $\alpha - n = \alpha$. If α be a non-limiting number, it is however convenient to denote the number immediately preceding α by $[\alpha - 1]$.

For
$$n + \omega = \omega,$$
since
$$\begin{aligned} n + \omega &= \overline{(e_1, e_2, \ldots e_n; f_1, f_2, \ldots f_n, \ldots)} \\ &= \overline{(g_1, g_2, \ldots g_n, g_{n+1}, \ldots)}, \end{aligned}$$
where
$$g_1 = e_1,\ g_2 = e_2, \ldots\ g_n = e_n,\ g_{n+1} = f_1,\ g_{n+2} = f_2, \ldots.$$

Further,
$$n + \alpha = n + \omega + (\alpha - \omega) = \omega + (\alpha - \omega) = \alpha.$$

If n is a finite number, then $n\omega = \omega$. This is seen by taking an aggregate of the type ω, and replacing each element by n new elements; then it is clear that the new aggregate is also of type ω.

It can be easily proved that $(\alpha + n)\,\omega = \alpha\omega$, where α is of the second class, and n of the first class.

173. *If α is any number of the second class, then those numbers of the first and second classes which are less than α form a normally ordered aggregate of type α, when they are arranged in order as defined above.*

If M is an aggregate such that $\bar{M} = \alpha$, and if α' is an ordinal number $< \alpha$, then there is a segment M', of M, such that $\bar{M}' = \alpha'$; and, conversely, every segment of M determines a number of the first, or the second, class which is $< \alpha$. For, since $\bar{\bar{M}} = \aleph_0$, any segment M' must have either a finite cardinal number, or else must have $\aleph_0$ for its cardinal number. If e_1 is the lowest element of M, a segment M' is determined by an element $e' > e_1$; and every element e', of M, determines a segment M'. If e', e'' are two elements of M, both $> e_1$, and M', M'' the segments of M determined by these elements, and α', α'' their order-types, then if $e' < e''$, it follows, by § 167, that $M' < M''$; and hence $\alpha' < \alpha''$.

If then $M = (e_1, M')$, and to the element e', of M', we make the element α', of $\{\alpha'\}$, correspond, the two aggregates M' and $\{\alpha'\}$ are placed in the relation of similarity. It has thus been shewn that $\overline{\{\alpha'\}} = \bar{M}'$; now $\bar{M}' = \alpha - 1 = \alpha$, hence $\overline{\{\alpha'\}} = \alpha$.

Since $\bar{\alpha} = \aleph_0$, we have $\overline{\overline{\{\alpha'\}}} = \aleph_0$; and therefore the following theorem is established:

The aggregate $\{\alpha'\}$ of all those numbers α', of the first and second classes, which are ordinally smaller than a number α, of the second class, has the cardinal number $\aleph_0$.

174. *Every number α, of the second class, is either* (1), *such that it is obtained from a number of the same class immediately preceding it, by the addition of unity, or else* (2), *such that there exists a sequence $\{\alpha_n\}$, of numbers of the first or second class, having α for its limit.*

Let $\alpha = \bar{M}$; then, if M has a highest element e, $M = (M', e)$, where M' is the segment of M determined by e; thus $\bar{M} = \bar{M}' + 1$, or

$$\alpha = [\alpha - 1] + 1.$$

If M has no highest element, then the aggregate $\{\alpha'\}$, of all numbers $< \alpha$, which is similar to M, has no greatest number; and this aggregate $\{\alpha'\}$, being of cardinal number $\aleph_0$, can be re-arranged as an aggregate $\{\alpha'_n\}$ of type ω. In this aggregate $\{\alpha'_n\}$, some of the numbers α_2', α_3', ... will in general be less than α_1', but others must be greater than α_1'; for α_1' cannot be greater than all the other numbers of the aggregate, there being in $\{\alpha'\}$ no greatest number. Let α'_{p_2} be that number of $\{\alpha'_n\}$, with the smallest index, such that $\alpha'_{p_2} > \alpha_1'$; similarly let α'_{p_3} be that number, with the smallest index, such that $\alpha'_{p_3} > \alpha'_{p_2}$, and so on. We have now an infinite sequence

$$\alpha_1', \alpha'_{p_2}, \alpha'_{p_3}, \ldots$$

of numbers, such that they are in ascending order, and such that their indices are also in ascending order. Since $n \leq p_n$, we have $\alpha'_n \leq \alpha'_{p_n}$; hence, for every number α' which is less than α, there exists a number α'_{p_n} which is $> \alpha'$. Since α is the number which follows next after all the numbers α', it is also the number which follows next after all the numbers α_1', α'_{p_2}, α'_{p_3}, ..., which we may write as $\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n, \ldots$; thus

$$\alpha = \lim_{n\sim\infty} \alpha_n.$$

It has thus been shewn that there are two kinds of numbers of the second class, (1), those which have an immediate predecessor in the aggregate of all such numbers arranged in ascending order, and (2), those which have no such immediate predecessor, and are called *limiting numbers.*

A number of the first kind is obtained by means of the first principle of generation, (see § 65), from the immediately preceding number.

A number of the second kind is obtained by the second principle of generation, as the number α which next follows all the numbers α_n, of some sequence $\{\alpha_n\}$ of numbers of the second class.

THE CARDINAL NUMBER OF THE SECOND CLASS OF ORDINALS

175. *The totality of the numbers of the second class, arranged in ascending order, forms a normally ordered aggregate.*

If A_α denotes the ordered aggregate of all those numbers of the second class which are less than the given number α, then A_α is normally ordered, and of type $\alpha - \omega$. For the aggregate $\{\alpha'\}$ of numbers of the first and second classes, which consists of $\{n\}$ and A_α, has been shewn, in § 173, to be normally ordered, and thus

$$\{\alpha'\} = (\{n\}, A_\alpha),$$

hence $$\overline{\{\alpha'\}} = \overline{\{n\}} + \bar{A}_\alpha, \text{ or } \bar{A}_\alpha = \alpha - \omega.$$

Let M denote any part of the aggregate $\{\alpha\}$, of all the numbers of the second class, such that, in $\{\alpha\}$, there are numbers which are greater than all the numbers in M; and let α_0 be one such number. then M is a part of A_{α_0+1}, which is such that all the numbers of M are less than at least one number α_0, of A_{α_0+1}. Since A_{α_0+1} is normally ordered, there must be a number α', of A_{α_0+1}, being itself consequently a number of $\{\alpha\}$, which is the next greater number than all the numbers of M. Thus, since $\{\alpha\}$ has a lowest number ω, the conditions are satisfied that $\{\alpha\}$ is a normally ordered aggregate.

It follows, by applying the results of § 165, that:

Every part of the aggregate $\{\alpha\}$, of all numbers of the second class, has a least number.

Every such part, in order, is normally ordered.

It will now be shewn that *the aggregate $\{\alpha\}$, of all the numbers of the second class, has a cardinal number greater than $\aleph_0$.*

Consider the two aggregates $\{\alpha\}$ and $\{n\}$, where n denotes the finite integers. The aggregate $\{\alpha\}$ has a part, viz. that consisting of

$$\omega + 1, \omega + 2, \dots \omega + n, \dots,$$

that is equivalent to $\{n\}$. It has been shewn in § 151 that every part of $\{n\}$ is either finite, or that it is equivalent to $\{n\}$ itself. In order to shew that the cardinal number of $\{\alpha\}$ is greater than $\aleph_0$, it is therefore only necessary to shew that it is not equal to $\aleph_0$.

If $\overline{\overline{\{\alpha\}}} = \aleph_0$, the numbers of $\{\alpha\}$ could be arranged in the form

$$\gamma_1, \gamma_2, \cdots \gamma_n, \cdots,$$

of type ω, in which of course the order would not be that of generation. Starting from γ_1, let γ_{p_2} be the γ, with the smallest index, which is such that $\gamma_{p_2} > \gamma_1$; then let γ_{p_3} be that γ, with the smallest index, such that $\gamma_{p_3} > \gamma_{p_2}$; and so on. We obtain in this manner a sequence

$$\gamma_1, \gamma_{p_2}, \gamma_{p_3}, \cdots$$

in ascending order, the indices $1, p_2, p_3, \ldots$ being also in ascending order. In accordance with § 171, there must be a definite number δ of the second class, namely $\delta = \lim_{n \sim \infty} \gamma_{p_n}$, such that $\delta > \gamma_{p_n}$, for every p_n, and consequently such that δ is greater than every γ_n, but this is impossible, since $\{\gamma_n\}$ contains every number of the second class; hence $\overline{\overline{\{\alpha\}}}$ cannot equal $\aleph_0$, and is therefore $> \aleph_0$.

Every part of the aggregate $\{\alpha\}$, of all numbers of the second class, has either the cardinal number of $\{\alpha\}$, or else the cardinal number $\aleph_0$, unless it is a finite part.

Every such part, when the elements of it are in order of generation, being part of the normally ordered aggregate $\{\alpha\}$, is either similar to $\{\alpha\}$, or else to some segment A_{α_0}, of $\{\alpha\}$; hence the cardinal number is either that of $\{\alpha\}$, or is $\overline{\overline{A_{\alpha_0}}} = \overline{\alpha_0 - \omega}$, and this last is either $\aleph_0$, or is finite.

The cardinal number of $\{\alpha\}$ is the cardinal number next greater than $\aleph_0$.

If there exist a cardinal number less than $\overline{\overline{\{\alpha\}}}$, and greater than $\aleph_0$, it must be the cardinal number of some part of $\{\alpha\}$; but it has been shewn that every such part of $\{\alpha\}$ has either the cardinal number of $\{\alpha\}$, or is $\aleph_0$, or is finite.

The cardinal number of $\{\alpha\}$, or of $Z\,\{\aleph_0\}$, is denoted by $\aleph_1$.

THE GENERAL THEORY OF ALEPH-NUMBERS

176. It has now been shewn that the ordinal numbers of the second class, in their order of generation, form a normally ordered aggregate, of which the cardinal number is $\aleph_1$, the next cardinal number to $\aleph_0$. The ordinal type of the normally ordered aggregate $\{\alpha\}$ of all numbers of the second class, is a number Ω, which is the smallest number of the third class. In analogy with the definition of the second class, and in accordance with what Cantor has denominated the principle of limitation (Hemmungsprinzip), the third class is taken to include all the ordinal types of normally ordered aggregates, of which the cardinal number is $\aleph_1$, and this class is consequently denoted by $Z\,(\aleph_1)$. The number Ω, which is the order-type of all the numbers of the first and second classes, in the order of generation, and which comes after all those numbers, is not the limiting element of any sequence $\alpha_1, \alpha_2, \dots \alpha_n, \dots$ of numbers of the second class; for, as we have seen, every such sequence has a limiting number within the second class. From the point of view adopted by Cantor in his earlier writings, and explained in § 65, in which the successive ordinal numbers are regarded as successively generated, in accordance with postulated principles of generation, the number Ω must be regarded as generated by a third principle of generation, different from the two principles of generation employed in the case of the numbers of the first and second classes. This third principle of generation affirms that every set of ordinal numbers similar to the aggregate of all the numbers of the first and second classes, in their order of generation, is immediately succeeded by a new number, ordinally greater than all the numbers of the set, so that every number which is less than this new number is also less than some of the numbers of the set. When, proceeding from Ω, the numbers

$$\Omega + 1, \Omega + 2, \dots \Omega + \Omega, \dots$$

are formed, all three principles of generation will be required, in forming the numbers of the third class.

From the point of view adopted later by Cantor, and explained in the present chapter, Ω is simply defined to be the order-type of the totality of the numbers of the first and second classes, in their normal order. The numbers higher than Ω are then defined in the same manner, each one as the order-type of the totality of the preceding numbers in normal order.

The existence of a whole series of classes of order-types of normally ordered aggregates, *i.e.* of ordinal numbers, has been speculatively asserted by Cantor[*], who has however, in his published works, confined his detailed investigations to numbers of the first and second classes.

To each of the successive classes of numbers, there corresponds a single cardinal number, that of the totality of the ordinal numbers up to, and including all the ordinal numbers of that class. The first ordinal number of each class is the order-type of all the numbers of the preceding classes, in their order of generation. A new principle of generation is required for the first number of each new class, since that number cannot be regarded as the limiting number of any sequence of which the ordinal number is less than that of the number in question. All the successive principles of generation are however included in the one principle, that an aggregate of normally ordered ordinal numbers has itself an order-type which is a new number; and thus, from this point of view, all the principles of generation, from the second, onwards, are replaced by this one principle.

---

[*] See *Math Annalen*, vol. XXI (1883), pp. 587, 588, also vol. XLVI (1895), p. 495.

177. In accordance with this theory, there exists an ordered aggregate

$$1, 2, 3, \dots n, \dots \omega, \omega + 1, \dots \omega^2, \dots \omega^\omega, \dots \Omega, \Omega + 1, \dots \gamma, \dots$$

which contains every ordinal number of every class; and there also exists a similar aggregate

$$1, 2, 3, \dots n, \dots \aleph_0, \aleph_1, \aleph_2, \dots \aleph_n, \dots \aleph_\omega, \dots \aleph_\Omega, \dots \aleph_\gamma, \dots$$

of cardinal numbers, each element of which is the cardinal number of a single class of numbers of the first aggregate.

That the first of these aggregates is normally ordered may be seen by remarking that, if it contained any part, of the type $^*\omega$, then such part would also be part of the normally ordered aggregate formed by the numbers $1, 2, 3, \dots \omega, \dots \alpha$; where α is the highest number in the hypothetical part, of type $^*\omega$. This is impossible, and hence the first aggregate is normally ordered.

Cantor has given a proof that $\aleph_0$ is less than, or equal to, the cardinal number of any transfinite aggregate, and that $\aleph_1$ is the cardinal number next greater than $\aleph_0$. In his proof the possibility of two cardinal numbers being incomparable was disregarded. Accordingly, the first theorem in § 151 has been made to precede Cantor's theorem, which has however also been given. A proof has been given by Jourdain†, that $\aleph_2$ is the next greater cardinal number than $\aleph_1$, who has also considered, in some detail, the ordinal numbers of the third class, and has given indications of extension to the higher classes.

The question whether every transfinite cardinal number is necessarily an Aleph-number, which is equivalent to the question whether every aggregate is capable of being normally ordered, has engaged a considerable amount of attention. That the answer should be an affirmative one, was regarded by Cantor as probable. Some discussion of attempts which have been made to settle this matter, will be considered in § 203. A case of great importance is that of the continuum, which is defined as a simply ordered, but not as a normally ordered, aggregate. No proof has yet been discovered of the correctness of Cantor's view that $c = \aleph_1$. In case c occurs at all in the aggregate of Aleph-numbers, the continuum is capable of being normally ordered. The possibility has also been contemplated that c may be greater than all the Aleph-numbers.

THE ARITHMETIC OF ORDINAL NUMBERS OF THE SECOND CLASS

178. The ordinal numbers of the second class have been defined as the order-types of normally ordered, enumerably infinite, aggregates; and the operations of addition and multiplication have been defined for these numbers, in §§ 169 and 170. It now remains for us to define exponentials, for numbers of this class; and the definition is founded upon the following theorem:

---

† See *Phil. Mag.* (6), vol. VII (1904), p. 294, "On the transfinite cardinal numbers of number-classes in general."

If ξ is a variable of which the domain consists of the numbers of the first and second classes, including zero, and if γ, δ denote two constants belonging to the same domain, such that $\delta > 0$, $\gamma > 1$, then there exists a single-valued determinate function $f(\xi)$, which satisfies the conditions

(1). $f(0) = \delta$.

(2). If ξ', ξ'' are any two values of ξ such that $\xi' < \xi''$, then $f(\xi') < f(\xi'')$.

(3). For every value of ξ, $f(\xi + 1) = f(\xi) . \gamma$.

(4). If $\{\xi_n\}$ is a sequence, of which ξ is the limiting number, then $\{f(\xi_n)\}$ is a sequence, of which $f(\xi)$ is the limiting number.

In the case $\delta = 1$, the function $f(\xi)$ is denoted by γ^ξ; and then $f(\xi)$, satisfying the above conditions, defines the exponential function γ^ξ, for all numbers γ, ξ of the first and second classes.

To prove the theorem, in the first place we have

$$f(1) = \delta\gamma,\ f(2) = \delta\gamma^2, \ldots,\ f(n) = \delta\gamma^n:$$

thus $f(1) < f(2) < f(3), \ldots$; and the function is determined for every $\xi < \omega$. Next assume that the function is determined for every $\xi < \alpha$, a number of the second class. If α is not a limiting number,

$$f(\alpha) = f[\alpha - 1]\gamma > f[\alpha - 1]:$$

and thus $f(\alpha)$ is determined. If α is a limiting number, and is preceded by the sequence $\{\alpha_n\}$, then $\{f(\xi_n)\}$ is a sequence, and $f(\alpha) = \lim\limits_{n\sim\infty} f(\alpha_n)$. If $\{\alpha_n'\}$ is another sequence such that $\alpha = \lim\limits_{n\sim\infty} \alpha_n'$, then the two sequences $\{f(\alpha_n)\}$, $\{f(\alpha_n')\}$ are related to one another, and therefore have the same limit; and thus $f(\alpha)$ is uniquely determined. $f(\xi)$ is now determined for every ξ: for if there were values of α for which it were not determined, there must be a smallest of such values; the theorem would then hold for $\xi < \alpha$, but not for $\xi \geqq \alpha$; which is contrary to what has been proved above.

179. If α, β are numbers of the first or second class, $\gamma^{\alpha+\beta} = \gamma^\alpha . \gamma^\beta$.

The function $\phi(\xi)$, $= \gamma^{\alpha+\xi}$, satisfies the conditions

(1), $\phi(0) = \gamma^\alpha$.

(2), If $\xi' < \xi''$, $\phi(\xi') < \phi(\xi'')$.

(3), $\phi(\xi + 1) = \phi(\xi) . \gamma$.

(4), If $\{\xi_n\}$ is a sequence such that $\lim\limits_{n\sim\infty} \xi_n = \xi$, then $\phi(\xi) = \lim\limits_{n\sim\infty} \phi(\xi_n)$.

It follows by § 178 that, if we take $\delta = \gamma^\alpha$, then $\phi(\xi) = \gamma^\alpha\gamma^\xi$; hence if $\xi = \beta$, we have

$$\gamma^{\alpha+\beta} = \gamma^\alpha . \gamma^\beta.$$

Again, if α, β are two numbers of the first or second class, If we put $f(\xi) = \gamma^{\alpha\xi}$, we find by applying the theorem of the last section, that $f(\xi) = (\gamma^\alpha)^\xi$, where γ^α replaces γ.

γ being > 2, it can be proved that, for every ξ, $\gamma^\xi \geq \xi$. The theorem holds for $\xi = 0$, $\xi = 1$; and if it be assumed to hold for all values of ξ which are less than a given number α, then it holds also for $\xi = \alpha$.

For, first let α be not a limiting number: then, if $[\alpha - 1] \leq \gamma^{[\alpha-1]}$, we have

$$[\alpha - 1]\gamma \leqq \gamma^\alpha;$$

hence $\gamma^\alpha \geqq [\alpha - 1] + [\alpha - 1][\gamma - 1]$. therefore since $[\alpha - 1]$ and $[\gamma - 1]$ are > 1, and $[\alpha - 1] + 1 = \alpha$, we have $\gamma^\alpha \geq \alpha$. If α is a limiting number $= \lim\limits_{n\sim\infty} \alpha_n$, then since $\alpha_n \leqq \gamma^{\alpha_n}$ we have $\lim\limits_{n\sim\infty} \alpha_n \leqq \lim\limits_{n\sim\infty} \gamma^{\alpha_n}$, or $\alpha \leq \gamma^\alpha$. If there were values of ξ such that $\xi > \gamma^\xi$, there would be one of such values which is the least of all; and if this were α, then $\xi \leqq \gamma^\xi$, if $\xi < \alpha$, but $\alpha > \gamma^\alpha$; which is contrary to what has been proved above.

180. Of all the numbers of the second class, the smallest ones are those which are algebraical functions of ω, of the form

$$\omega^n.p_n + \omega^{n-1}.p_{n-1} + \dots + \omega.p_1 + p_0,$$

where $p_0, p_1, \dots p_n$ are finite numbers. If we write

$$\omega_1 = \omega^\omega, \quad \omega_2 = \omega^{\omega_1}, \quad \omega_3 = \omega^{\omega_2}, \dots$$

then we obtain the number $\epsilon_0 = \lim\limits_{n\sim\infty} \omega_n$. This number ϵ_0 is the smallest of a species of numbers of the second class which are characterized by the property $\epsilon = \omega^\epsilon$, and which Cantor has designated ϵ-numbers. Cantor has shewn that the ϵ-numbers form a normally ordered aggregate, of type Ω, and therefore similar to the whole second class of numbers. He has further shewn that every number α, of the second class, is uniquely representable in the form

$$\alpha = \omega^{\alpha_0}\kappa_0 + \omega^{\alpha_1}\kappa_1 + \dots + \omega^{\alpha_r}\kappa_r,$$

where $\alpha_0, \alpha_1, \alpha_2, \dots \alpha_r$ are numbers of the first or second class which satisfy the conditions $\alpha_0 > \alpha_1 > \alpha_2 > \dots > \alpha_r \geqq 0$, and $\kappa_0, \kappa_1, \kappa_2, \dots \kappa_r, r + 1$, are numbers of the first class which are different from zero. For the detailed investigation of the normal form, and for that of the special class of ϵ-numbers, we must refer to Cantor's original discussion*.

THE THEORY OF ORDER-FUNCTIONS

181. A method of representation of any mode of ordering a given aggregate M has been given by Bernstein†. When the elements of the aggregate are numbers, this method lends itself to a diagrammatic representation of the aggregate, as ordered in any particular order-type.

---

* *Math. Annalen*, vol. XLIX (1897), pp. 235–246.

† See his Dissertation, also W. H. Young, on "Closed sets of points and Cantor's numbers," *Proc. Lond. Math. Soc.* (2), vol. I (1903), p. 230.

An aggregate M is ordered, in the most general sense of the term, when it is known as regards every pair of elements a, b, whether $a \gtrless b$; but a particular mode of ordering the aggregate can be represented by means of a function $f(a, b)$ of the pairs of elements, which is defined by

$$f(a, b) = 1, \text{ if } a < b; f(a, b) = -1, \text{ if } a > b; \text{ and } f(a, a) = 0.$$

This function $f(a, b)$ may be denominated an order-function of the aggregate M; and there is one order-function for each possible mode of ordering the aggregate. The function must satisfy the condition that, if

$$f(a, b) = f(b, c),$$

then each is equal to $f(a, c)$.

Two order-functions $f_1(a, b)$, $f_2(a, b)$, of a given aggregate M, represent two methods of ordering the aggregate in one and the same order-type, provided there exist a reversible transformation ϕ, of the aggregate M into itself, such that $f_1\{\phi(a), \phi(b)\} = f_2(a, b)$.

All those order-functions of a given aggregate M, which correspond to an arrangement of M in one and the same order-type, constitute a family of order-functions; and there is one such family of order-functions corresponding to each order-type in which the given aggregate M can be arranged.

It is clear that the order-functions of a family corresponding to $\bar{M}$ form an aggregate with the same cardinal number as the group of transformations of $\bar{M}$ into itself.

If, in particular, the aggregate M is that of the positive integers, then a pair of elements (a, b) is represented by a cross-point of the rectangular trellis formed in the positive quadrant by drawing all the straight lines $x = i$, $y = i$, for positive integral values of i, referred to rectangular Cartesian coordinates x, y.

The natural order of the numbers 1, 2, 3, ... will be represented by $f(x, y)$, defined for all the cross-points, so that $f(x, y) = 1$, when $x < y$, and $f(x, y) = -1$, when $x > y$, and such that $f(x, x) = 0$.

Any particular mode* of ordering the numbers 1, 2, 3, ... will be represented by marking one set of cross-points $+1$, and another set -1, those on the diagonal $x = y$ being marked zero.

It is, however, not every mode of so marking the cross-points that represents a possible ordering of the aggregate. That a mode of marking may represent a possible order, two conditions must be satisfied. First, we must have $f(x, y) = -f(y, x)$; and thus points which are optical images, relatively to the diagonal $y = x$, must be marked with unities of opposite sign. Secondly, the condition that, if $f(a, b_1) = f(b_1, c)$, then each $= f(a, c)$, must be satisfied. This condition may be expressed as follows:—Join every cross-point which is marked $+1$, or 0, with every other such cross-point, then the resulting figure may be called the positive frame-work; join similarly all the pairs of points marked $-1, 0$; in this way we obtain the negative frame-work. Let lines joining the two pairs of points (x_1, y_1), (x_2, y_2) and (x_1, y_2), (x_2, y_1) be called conjugate lines. The condition which must be satisfied is that no side of the positive frame-work can be conjugate to a side of the negative frame-work. It can easily be seen that the condition so stated is necessary and sufficient.

---

* Some examples of order-types represented in this manner are given by W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. I (1903), p. 244.

THE CARDINAL NUMBER OF THE CONTINUUM

182. The arithmetic continuum has been defined as an aggregate of the order-type θ (see § 163), and it is thus not normally ordered. It has been held by Cantor* that this aggregate, and perhaps every aggregate, is capable of being arranged as a normally ordered aggregate; but no universally accepted proof of the correctness of this view has been obtained. If the continuum be capable of arrangement as a normally ordered aggregate, its cardinal number c must be identical with one of the Aleph-numbers; and in fact Cantor believed that $c = \aleph_1$, the cardinal number of the aggregate of all the order-types of normally ordered enumerable aggregates. As evidence of the probable truth of this view, the facts may be cited that all the sets of points which have actually been defined in connection with the theory of sets of points have one or other of the two cardinal numbers $\aleph_0$ and c, and that no such set of points has been defined of which it is known that the cardinal number is $> \aleph_0$ and $< c$. This negative evidence is however clearly insufficient to settle the question whether every part of the continuum has one of the powers $\aleph_0$ or c, a question which has hitherto defied all attempts to obtain a conclusive answer. It has been shewn by König that the cardinal number of the continuum cannot be an Aleph-number of which the index is a limiting number. As has already been pointed out, it cannot be assumed that every two cardinal numbers are such as to be comparable with one another; but a proof has been propounded by G. H. Hardy† that c, and presumably any cardinal number whatever, must either be an Aleph-number, or else be greater than all the Aleph-numbers. The mode of reasoning is a generalization of that employed by Cantor in his proof (see § 151) that $\aleph_0$ is less than any other transfinite cardinal number.

Because $c \geqq \aleph_1$, it is possible to take elements from the number-continuum corresponding to all the numbers of the first and second classes of ordinals. For, if this process came to an end, we should have $c = \aleph_0$, which has been proved by Cantor not to be the case. It follows that a set can be selected from the continuum equivalent to the aggregate of ordinal numbers of the first and second classes. Now if a set could be selected from this aggregate equivalent to the continuum, it would follow from the equivalence theorem, proved in § 153, that $c = \aleph_1$; and if no such set could be selected it would follow from the definition of inequality in § 146, that $c > \aleph_1$; thus it follows that $c \geqq \aleph_1$. If now $c > \aleph_1$, a similar proof would shew that $c \geqq \aleph_2$, and so on. If $c > \aleph_n$ for every finite n, then $c \geqq \aleph_\omega$, and this process may be continued indefinitely through the Aleph series. The validity of this proof depends upon considerations which will be discussed in § 198

---

* *Math. Annalen*, vol. XXI (1883), p 550.

† *Quarterly J. of Math.* vol. XXXV (1903), p. 87.

It is known that every infinite closed linear set of points has one or other of the two cardinal numbers $\aleph_0$, c; and if a set of points can exist of which the cardinal number has neither of these two values, it must be unclosed, and may without loss of generality be taken as dense in itself. The difficulties of dealing with unclosed sets, dense in themselves, are so great that attempts to find a contradiction involved in the assumption of the existence of such a set, possessing a cardinal number different from both $\aleph_0$ and c, have hitherto been a complete failure.

183. A very remarkable relation has been given by Cantor between the cardinal number of the continuum and that of the integral numbers This relation is expressed by $c = 2^{\aleph_0}$, or more generally $c = n^{\aleph_0}$, where n is a finite integer.

This theorem was applied by its discoverer to obtain a simple arithmetical proof that the $\aleph_0$-dimensional continuum has the same power as the one-dimensional continuum.

In accordance with the definition of an exponential, given in § 150, $2^{\aleph_0}$ is the cardinal number of the numbers in the dyad scale,

$$\frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \dots;$$

where every b is either 0 or 1. In this aggregate, each number of the form $\frac{2p+1}{2^q} < 1$, where p and q are integers, occurs twice; hence

$$2^{\aleph_0} = (\overline{\overline{s, X}}),$$

where X is the aggregate of real numbers between 0 and 1, and s is an enumerable aggregate.

It follows from the above, that $2^{\aleph_0} = c + \aleph_0$. Now $c + \aleph_0 = c + 2\aleph_0$, since $\aleph_0 = 2\aleph_0$; therefore $c = c + \aleph_0$; whence we have $2^{\aleph_0} = c$.

From this theorem, we deduce $c \, . \, c = 2^{\aleph_0} . 2^{\aleph_0} = 2^{2\aleph_0} = 2^{\aleph_0} = c$; and hence by repeated multiplication by c, we find $c^n = c$, where n is any finite integer.

Again $c^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 . \aleph_0} = 2^{\aleph_0} = c$,
and therefore the continuum of finite, or of enumerable, dimensions is equivalent to the one-dimensional continuum.

The aggregate defined by all possible modes of distributing the numbers of the continuum upon themselves, has the power $c^c \equiv f$; and this number f is greater than c. This has been proved in § 154; for a part of the new aggregate is equivalent to that obtained by replacing the numbers of the continuum either by A, or by B, and taking all possible aggregates which arise in this way. Since this part has been shewn to have a cardinal number greater than that of the original aggregate, it follows then that $f > c$.

More generally, if a is any cardinal number, we have $a^a > a$.

*If the continuum be divided into any finite number n, of parts, such that all the parts have the same cardinal number, then that cardinal number is the same as that of the continuum**.

The parts may consist of sets of points of any kind.

The theorem may also be stated thus:—if $na = c$, then $a = c$.

To prove it, we have $na = c = nc$; and therefore, by applying the theorem of § 156, it follows that $a = c$.

184. *The continuum is equivalent to the aggregate of all possible order-types of simply ordered aggregates of cardinal number $\aleph_0$.*

This theorem points the contrast between the aggregate of all order-types of simply ordered enumerable aggregates, which is of power $2^{\aleph_0}$, and the aggregate of all the order-types of normally ordered enumerable aggregates, which is $\aleph_1$. The latter aggregate is, of course, a part of the former one, and thus the theorem $\aleph_1 \leq 2^{\aleph_0}$ can be deduced.

The theorem may be also stated in the form: *The total number of ways of ordering the integral numbers* 1, 2, 3, ... *is c.*

If μ is the order-type of an enumerable aggregate, arranged as a simply ordered aggregate, then a part of an aggregate of the type η, considered in § 162, can always be determined, which is of the type μ. To establish this, it can be shewn that an aggregate of type μ can always be changed into one of type η, by insertion of new elements. If, between every pair of elements m_1, m_2, of μ, there are other elements, then μ is of one of the types η, $1 + \eta$, $\eta + 1$, $1 + \eta + 1$; so that μ is reduced to η by the removal of the lowest and the highest elements, when such elements exist. In any such case, if we add to μ aggregates of type η, at the beginning and at the end, we obtain an aggregate of type $\eta + \mu + \eta$, which is of type η, whichever of the types η, $1 + \eta$, $\eta + 1$, $1 + \eta + 1$ may be identical with μ.

---

* Bernstein, *loc. cit.* p. 31.

If pairs of elements exist in the aggregate of type μ, such that there are no elements between them, an aggregate of type η can be inserted between every such pair, until a new aggregate of one of the types η, $1 + \eta$, $\eta + 1$, $1 + \eta + 1$, is obtained; then, as before, by adding aggregates of type η, at the beginning and end, we obtain an aggregate of type η. It has thus been shewn that any aggregate of type μ is a part of another aggregate of type η. Since the rational numbers in their totality naturally exist in the order-type η, it follows that an aggregate of any type μ can be made by taking a part of the aggregate of rational numbers, of type η. It follows that the aggregate of all types μ has a cardinal number less than, or equal to, that of the aggregate of all part-aggregates of the set of rational numbers, arranged in type η.

Now every aggregate $(r_1, r_2, \ldots)$, of which all the elements are rational numbers, corresponds to a single point of a continuum of an enumerable number of dimensions, of which the coordinates are $x_1 = r_1$, $x_2 = r_2, \ldots$. Hence the cardinal number of the aggregate of all part-aggregates of the set of rational numbers is less than, or equal to, the cardinal number of the $\aleph_0$-dimensional continuum, that is, $\leq c$; and therefore the cardinal number of the aggregate of all types μ is $\leq c$.

It will now be proved that $c \leq$ the cardinal number of the aggregate of all types μ. To every real number between 0 and 1, there corresponds an infinite sequence $b_1 b_2 b_3 \ldots$, where every b is either 0 or 1, expressing the number in the dyad scale. After each b, insert an aggregate of type π, and we then have an aggregate $b_1 \pi b_2 \pi b_3 \pi \ldots$, of type

$$\nu = b_1 + \pi + b_2 + \pi + b_3 + \pi + \ldots.$$

Here, some of the b's may be zero, and these may be simply omitted; thus $\pi + 0 + \pi = \pi + \pi$. Hence, to any real number x, between 0 and 1, there corresponds the type

$$\nu = b_1 + \pi + b_2 + \ldots.$$

It is now necessary to shew that the two order-types ν, ν', which correspond to two different numbers, x, x', are necessarily distinct from one another. If $\nu = \nu'$, we can write the equality

$$C_1 + \pi + \zeta_1 = C_1' + \pi + \zeta_1',$$

where C_1, C_1' are each either 0 or 1; and from this we obtain, by means of the theorem of § 164, $C_1 = C_1'$, and $\zeta_1 = \zeta_1'$. The last equation can be written

$$C_2 + \pi + \zeta_2 = C_2' + \pi + \zeta_2',$$

and from this we conclude that $C_2 = C_2'$, $\zeta_2 = \zeta_2'$; and we can proceed onwards in the same manner. From $b_1 = b_1'$, $b_2 = b_2', \ldots$, we conclude that $x = x'$. It has thus been shewn that $\{x\} \simeq \{\nu\}$; and from this we conclude that $c \leq$ the cardinal number of all order-types μ.

This part of the theorem is due to Cantor*, and the first part to Bernstein. By combining the two results, the complete theorem is established.

This important result may also be expressed by saying that *the totality of all permutations of the sequence of positive integers has the power of the continuum.*

It may also be shewn that *the totality of all parts of the sequence* 1, 2, 3, ... *has the power of the continuum.*

For, if we form a sequence, by writing 0 for each of the numbers 1, 2, 3, ... which does not occur in a given part of (1, 2, 3, ...); and 1 for each number which does occur in the given part, then the sequence of 0's and 1's, thus obtained, corresponds to a real number expressed in the dyad scale, and therefore the numbers of the continuum are put into correspondence with the parts of the sequence (1, 2, 3, ...).

It was thought probable by Cantor that $2^{\aleph_0} = \aleph_1$, but no proof of the correctness of this hypothesis has been obtained. The hypothesis has been discussed† by Sierpiński.

185. It can be shewn that *the aggregate of all sets of points in the n-dimensional continuum has a cardinal number greater than c.*

For, in the aggregate $\{P\}$ of all points in an n-dimensional continuum, we can substitute 0 for each point P which does not occur in a given set of points of the continuum, and 1 for each point P which does occur; we then obtain an aggregate consisting of 0's and 1's: but it is known that the totality of all such aggregates has the power 2^c, which is $> c$.

On the other hand, *the totality of all closed sets of points in the n-dimensional continuum has the same power c as the continuum.*

Every closed set is the derivative of an enumerable set of points; and, to every enumerable set of points, there corresponds a single closed set.

It follows that the cardinal number of the totality of closed sets is $\leq$ the cardinal number of the totality of enumerable sets of points chosen out of the continuum. To shew that the latter is c, we observe that it is $\leq$ the aggregate of all combinations of points of the continuum in sets of $\aleph_0$ elements, that is, $\leq c^{\aleph_0}$, or $\leq c$.

Again, every single point of the $\aleph_0$-dimensional continuum corresponds to a single point of the one-dimensional continuum, and this point is an enumerable part of the continuum; hence the totality of enumerable sets of points of the n-dimensional continuum is $\geq c$. On combining this with what has been proved above, we see that the totality of all enumerable sets of points in the n-dimensional continuum is c; hence the totality of all closed sets of points in the n-dimensional continuum is $\leq c$.

---

[*] See Bernstein's Dissertation, p. 7.

[†] *Fundamenta Math.* vol. v (1924), p. 177.

Again, the totality of all closed sets of points in the n-dimensional continuum is $\geq c$. For one such closed set can be taken in each of an infinity of the domains $x_1 = \alpha$, where x_1 is one of the n coordinates which determine the position of a point in the n-dimensional continuum; and the aggregate of all possible values of α has the power c. We thus obtain an aggregate of closed sets which has the power c; and it follows that the aggregate of all closed sets in the n-dimensional continuum is $\geq c$.

Since the totality of all closed sets of points in the n-dimensional continuum is $\geq c$, and at the same time is $\leq c$, it must have the cardinal number c.

Since* every curve, or surface, in a continuum is formed by a closed set of points, we see that every possible curve, or surface, corresponds uniquely to a single definite real number.

186. A method of constructing a set of points of which the cardinal number is $\aleph_1$, has been given† by G. H. Hardy.

If we start from the sequence

$$1,\ 2,\ 3,\ 4,\ 5,\ \ldots\ldots \quad (1)$$

of integral numbers, a new sequence

$$2,\ 3,\ 4,\ 5,\ \ldots\ldots \quad (2)$$

is formed by omitting the first term.

Continuing this process, we form

$$\begin{array}{l} 3,\ 4,\ 5,\ 6,\ \ldots\ldots \quad (3) \\ 4,\ 5,\ 6,\ 7,\ \ldots\ldots \quad (4) \\ 5,\ 6,\ 7,\ 8,\ \ldots\ldots \quad (5) \\ \ldots\ldots\ldots\ldots \end{array}$$

We now form a new sequence

$$1,\ 3,\ 5,\ 7,\ 9,\ \ldots\ldots \quad (\omega)$$

by traversing the above infinite array of sequences diagonally. Then we form

$$\begin{array}{l} 3,\ 5,\ 7,\ 9,\ 11,\ \ldots\ldots \quad (\omega + 1) \\ 5,\ 7,\ 9,\ 11,\ 13,\ \ldots\ldots \quad (\omega + 2) \\ 7,\ 9,\ 11,\ 13,\ 15,\ \ldots\ldots \quad (\omega + 3) \\ 9,\ 11,\ 13,\ 15,\ 17,\ \ldots\ldots \quad (\omega + 4) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 1,\ 5,\ 9,\ 13,\ 17,\ \ldots\ldots \quad (\omega . 2) \\ 5,\ 9,\ 13,\ 17,\ 21,\ \ldots\ldots \quad (\omega . 2 + 1) \\ 9,\ 13,\ 17,\ 21,\ 25,\ \ldots\ldots \quad (\omega . 2 + 3) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 1,\ 9,\ 17,\ 25,\ 33,\ \ldots\ldots \quad (\omega . 3) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}$$

Thus sequences corresponding to all the numbers $\omega . \mu + \nu$ can be formed.

---

* Bernstein, *loc. cit.* p. 43

† *Quarterly Journal of Math.* vol. xxxv, 1903, "A theorem concerning the infinite cardinal numbers." See also Hausdorff, *Leipz. Ber.* vol. LIX (1907), p. 155.

To form the sequences corresponding to ω^2, we take the array of sequences

$$\begin{array}{rrrrrl}
1, & 3, & 5, & 7, & 9, & \ldots\ldots\ (\omega) \\
1, & 5, & 9, & 13, & 17, & \ldots\ldots\ (\omega.2) \\
1, & 9, & 17, & 25, & 33, & \ldots\ldots\ (\omega.3) \\
1, & 17, & 33, & 49, & 65, & \ldots\ldots\ (\omega.4) \\
1, & 33, & 65, & 97, & 129, & \ldots\ldots\ (\omega.5) \\
\ldots & \ldots & \ldots & \ldots & \ldots &
\end{array}$$

and traverse it diagonally; we thus obtain

$$1,\ \ 5,\ 17,\ 49,\ 129,\ \ldots\ldots\ (\omega^2).$$

Generally, if $b_1, b_2, b_3, b_4, \ldots$ is the sequence corresponding to β, the sequence $b_2, b_3, b_4, b_5, \ldots$ corresponds to $\beta + 1$. To obtain a sequence corresponding to a number γ which is a limiting number of the second class, we take the array of sequences corresponding to any ascending set of numbers $\beta_1, \beta_2, \ldots$ of which the limit is γ, and traverse it diagonally. It is clear that, in this manner, a sequence can be found for any given number of the second class; but that the set of sequences so obtained is not unique. For example, ω might have been taken as the limit of $1, 3, 5, 7, \ldots$, or ω^2 might have been taken as the limit of

$$\omega + 1, \omega.2 + 2, \omega.3 + 3, \ldots.$$

It will be shewn that the sequences $b_1, b_2, b_3, \ldots$ can be so chosen that in every case $b_1 < b_2 < b_3 < \ldots$; and that, if the sequences $b_1, b_2, \ldots$ and $b_1', b_2', \ldots$ correspond to any two numbers β, β', where $\beta < \beta'$, then there exists a number N such that $b_n' > b_n$, for $n \geqq N$; and thus that the sequences are distinct from one another.

Let us assume that sequences, corresponding to all numbers $< \gamma$, have been constructed in such a manner that this condition is satisfied. First, let γ be a non-limiting number, so that $\gamma = \gamma' + 1$. Then if $\beta < \gamma'$, there is a number N such that $a_n' > b_n$, for $n \geqq N$, where $a_1', a_2', a_3', \ldots$ form the sequence which corresponds to γ'. But if $a_1, a_2, a_3, \ldots$ form the sequence which corresponds to γ, we have $a_n = a'_{n+1} > a_n' > b_n$, for $n \geqq N$.

Hence, if the construction is possible for all numbers $< \gamma$, it is possible for all numbers $\leqq \gamma$, where γ is a non-limiting number.

Next, let us suppose that γ has no immediate predecessor, and that $\gamma = \lim\limits_{m \sim \infty} \beta_m$; then also $\gamma = \lim\limits_{m \sim \infty} (\beta_m + \nu_m)$, where the ν_m are finite numbers. Now there is a number N_1, such that $b_{2,n} > b_{1,n}$, for $n \geqq N_1$, where $b_{m,n}$ denotes the nth number in the sequence corresponding to β_m. *A fortiori*, if $\gamma_m = \beta_m + \nu_m$, we have $c_{2,n} = b_{2,n+\nu_2} > b_{2,n} > b_{1,n}$; for $n \geqq N_1$, where $c_{m,n}$ is the nth number in the sequence corresponding to γ_m. But if we take $\nu_2 > b_{1,N_1-1}$, we have $c_{2,n} = b_{2,n+\nu_2} \geqq n + \nu_2 > b_{1,N_1-1} > b_{1,n}$, for $n < N_1$; and hence we have $c_{2,n} > b_{1,n}$, for all values of n. Similarly, ν_3 can be so chosen that $\gamma_3 > \gamma_2$, and $c_{3,n} > c_{2,n}$, for all values of n; and so on generally. If we write γ_1 for β_1, and $c_{1,n}$ for $b_{1,n}$, we have a doubly infinite array

$$\begin{array}{llll} c_{1,1}, & c_{1,2}, & c_{1,3}, & \ldots \\ c_{2,1}, & c_{2,2}, & c_{2,3}, & \ldots \\ c_{3,1}, & c_{3,2}, & c_{3,3}, & \ldots \end{array}$$

and we define the sequence corresponding to γ, by traversing it diagonally, so that $c_n = c_{n,n}$. If then $\beta < \gamma$, we can find m so that $\beta < \gamma_m$; then there is a number K, such that $c_{m,n} > b_n$, for $n \geq K$. But if $n > m$, we have $c_n = c_{n,n} > c_{m,n}$; and thus, if n is greater than the greater of the two numbers m, K, we have $c_n > b_n$. It has thus been shewn that, if the construction is possible for all numbers $< \gamma$, it is possible for all numbers $\leq \gamma$, whether γ is a limiting number or not.

In this manner, a sequence is obtained which corresponds to any assigned number γ of the second class, and this sequence is distinct from those which correspond to the numbers $< \gamma$, such sequences being also distinct from one another.

The sequences may be correlated with points in the linear continuum (0, 1). To correlate a sequence $b_1, b_2, b_3, \ldots$, we may take the binary radix fraction in which the b_1th, b_2th, b_3th, ... figures are all 1, and the remaining figures all 0. In this manner a set of points is shewn to exist, such that one point of the set corresponds to each number of the first, or of the second, class. This amounts to the construction of a set of points of cardinal number $\aleph_1$. Just as an enumerable set of points is determinate, when the point which corresponds to any assigned number n, of the first class, is determinate, so the set of cardinal number $\aleph_1$ is determinate, in the sense that a definite point is determined, corresponding to any assigned number β, of the first, or of the second, class.

It may be remarked that *a set of points of cardinal number $\aleph_1$, or of any cardinal number $> \aleph_0$, when arranged in normal order, cannot possibly be in the order in which they occur in the continuum.*

For, if a set of points, in the order in which they occur in the continuum, forms a normally ordered aggregate, each point and the next succeeding one define a linear interval of which they are the end-points. We have thus a set of intervals which must have the same cardinal number as the given set of points. Each interval of the set abuts on the next one, and thus the end-points together with their limiting points define an enumerable closed set. Hence the given set must be enumerable.

GENERAL DISCUSSION OF THE THEORY

187. Cantor's definition of an aggregate, given in § 145, implies that the elements of an aggregate are logically prior to the aggregate itself. The *definite* and *distinct* objects are, by an act of synthesis, regarded as a single collection, or aggregate. Such an act of synthesis involves a postulation, which is subject to the law of contradiction. It appears to have been fully recognized by Cantor himself, in connection with a certain so-called "inconsistent" aggregate, to be discussed below, that this is the case. The postulations of the existence, as single objects, of the aggregate of integral numbers, and of that of real numbers, are instances of such an act of synthesis. A somewhat different view of the nature of an aggregate is however widely spread. In accordance with this view, the elements of an aggregate consist of all objects that satisfy some prescribed condition or conditions; and thus the aggregate consists of a certain class of objects; and there has been a tendency to identify the conception of aggregate with that of "class," employed in Formal Logic. The aggregate itself is, from this point of view, logically prior to the elements it contains; in fact, in any given case, it may be doubtful whether any such elements exist; or it may even be known that no element exists, in which case the aggregate, or class, is called a null-aggregate, or null-class. Thus the relation of the aggregate to its elements is inverse to that which is in accordance with Cantor's synthetic definition. The following form of definition expresses this view of the nature of an aggregate:

All objects which are such as to satisfy a prescribed norm are said to belong to an aggregate defined by that norm. The norm consists of a set of specified conditions, or of a set of alternative specified conditions; and this norm must be sufficient to render it logically determinate, as regards any particular object whatever, whether that object belongs to the aggregate or not.

In the case of a finite aggregate the norm may take the form of individual specification of the objects which form the aggregate, but such exhaustive individual specification of the elements of a non-finite aggregate is not possible. The condition that it must be logically determinate whether a particular object is an element of a particular aggregate, or not, covers cases in which effective determination is not possible, either because such determination is impracticable, on account of the length of the process which would be requisite actually to make the decision, or because we are not in possession of the requisite means for making the decision. For example, the determination of the millionth digit in the value of the number π is merely impracticable, although it can be carried out by a finite process. Again, the question whether a particular number, such as Mascheroni's constant, belongs to the aggregate of transcendental numbers, has a logically determinate answer, although we may not be in possession of a method by which that answer can be obtained. The definition admits the logical determination as sufficient when effective determination is impossible or difficult. There may even be cases in which the effective determination of the question whether a defined aggregate is a null-aggregate, or not, is impossible, by the employment of existing means, or is so difficult as to be impracticable.

It is clear that the elements of an aggregate, being subject to a common norm, must have a certain community of nature, their class mark, which constitutes the ground of the aggregation

188. The discussion of the theory of aggregates which has taken place amongst mathematicians, since the publication of Cantor's theories, has led to the examination of certain aggregates which appear to satisfy, *prima facie*, the definition of an aggregate, and yet which are such that their existence, at all events if they are regarded as possessing cardinal numbers or order-types, involves certain contradictions. In this manner, what are known as the antinomies of the theory of aggregates have arisen. Both Cantor's definition of an aggregate, and the definition, given above, of an aggregate as a class, lead to antinomies of this kind, the most important of which will be discussed below. It was, however, recognized by Cantor* that the collection of elements into an aggregate is only valid provided a conception free from contradiction is the result of such synthesis. Suggestions have been made in various quarters, in the direction of limiting the concepts of the theory by means of postulates, in such a manner that the antinomous aggregates will not arise. At the present time however, no such generally accepted scheme of postulations and axioms is in existence. The final goal of such investigations would be attained if we were in possession of a definition of an aggregate, of such a character that it could be shewn *a priori* to be free from contradiction, without waiting for the discovery of antinomies. Such a definition should also be sufficiently wide; so that all aggregates, which are fitted for employment in a mathematical theory, would fall under it.

On systems of axioms which are suited for the building up of a theory of aggregates of such a character that no contradictions will arise, the writings of Hilbert†, Zermelo‡, Russell §, Schoenflies||, and Huntington¶ may be consulted. Developments of the theory which go beyond those

---

* See Jourdain, *Phil Mag.* (6), vol. VII (1904), p. 67.

† *Verhand. d. Math. Kongress z Heidelberg* (1904), p. 174; *Jahresber. d. Math. Vereinigung*, vol. VIII, 1 (1900), p. 180; *Math. Annalen*, vol XCV (1926), p. 161. See also Mollerup, *Math. Annalen*, vol LXIV (1907), p 231.

‡ *Math. Annalen*, vol. LXV (1908), p. 261; *Acta Math* vol. XXXII (1909), p. 186.

§ *Proc Lond. Math Soc* (2), vol. IV (1905), p. 29.

|| *Math. Annalen*, vol. LXXII (1912), p. 551, vol. LXXXIII (1921), p. 173, and vol. LXXXV (1922), p. 60.

¶ *Ann. of Math.* (2), vol. VI (1904), p. 151, and (2), vol. VII (1905), p. 15.

of which an account is given in the present chapter will be found in the writings of Hausdorff*, Hessenberg†, Schoenflies‡, Bernstein §, Jourdain|| and Whitehead¶. A treatise has been published by König**, in which the subject is treated from his own point of view.

An interesting discussion of the objections which have been raised in various quarters against Cantor's theory of aggregates has been given†† by Fraenkel, who has also developed in some detail the "axiomatic" treatment of the theory due to Zermelo, in which no direct definition of an aggregate, such as that of Cantor, is employed.

189. It is not clear that an aggregate, defined as a class of objects, is necessarily capable of being ordered at all. For example, it is difficult to see that such an aggregate as that of "all propositions" could conceivably be ordered; where it is assumed that the meaning of the word "proposition" is taken as so definite, that this aggregate has a norm in accordance with the definition above. Again, to take an example among aggregates of the kind usually considered in Mathematical theory, we may consider the aggregate obtained by distributing the aggregate of real numbers upon itself. This aggregate which has the cardinal number $f \equiv c^c$, is equivalent to the aggregate of all the functions of a real variable; it is difficult, if not impossible, to see how order could be effectively imposed upon this aggregate. If then, a transfinite aggregate is to be given as an ordered aggregate, or is to have an order imposed upon it, or rather discovered in it, it would appear to be necessary that the norm, which constitutes the definition of the aggregate, should be of such a character, that a principle of order is contained therein, or can at all events be adjoined thereto; so that, when any two particular elements are considered, the conditions which they satisfy in virtue of their belonging to the aggregate, when individualized for the particular elements, may be sufficient also to allow of relative rank being assigned to those elements, in accordance with a principle of order. This is in fact the case in such

---

* *Grundzuge der Mengenlehre,* Leipzig, 1914; also *Math. Annalen,* vol. lxv (1908), p. 435, and vol lxxv (1914), p. 428; *Leipz. Ber* vol. liii (1901), p 460; vol. lviii (1906), p. 106, and vol. lix (1907), p. 84; *Jahresber. d. Math. Vereinigung,* vol. xiii (1904), p. 570.

† *Grundbegriffe der Mengenlehre,* Göttingen, 1906.

‡ *Math. Annalen,* vol. lviii (1904), p. 195; vol. lix (1904), p. 129; vol. lx (1905), p. 181, and p. 431.

§ *Math. Annalen,* vol. lxi (1905), p. 117.

|| *Phil. Mag.* (6), vol. ix (1905), p. 42; (6), vol vi (1903), p. 323, and vol. vii (1904), pp 61 and 294.

¶ *Amer. J. of Math.* vol. xxiv (1902), p. 367, and *Archiv d. Math. u. Physik* (3), vol. x (1906), p. 254; also *Quarterly Journal,* vol. xxxix (1908), p. 375.

** *Neue Grundlagen der Logik, Arithmetik und Mengenlehre,* Leipzig (1914), also *Math. Annalen,* vol. lx (1905), p. 177, and p. 462. Also *Comptes Rendus,* vol. cxliii (1906), p. 110, and *Verhand d. Math. Kongress z. Heidelberg* (1904), p. 144.

†† See his work, *Einleitung in die Mengenlehre,* 2nd ed., Berlin (1923). See also Chwistek, *Math. Zeitschr.* vol. xxv (1926), p. 439.

aggregates as those of the integral numbers, the rational numbers, or the real numbers. In the case, for example, of the positive rational numbers, the relative rank of any two particular elements (p, q), (p', q') is assigned by the system of postulations, contained in § 12, which defines the aggregate. It may, of course, also be possible in other cases, as in this one, to re-order the aggregate, in accordance with some other law, extrinsically imposed upon the aggregate; but the nature of the elements must be such that this is possible.

We can now state that:

In order that a transfinite aggregate, defined as in § 187, *may be capable of being ordered, a principle of order must be explicitly or implicitly contained in the norm by which the aggregate is defined.*

The relative order of any two elements chosen from an ordered aggregate depends upon the individual characteristics of those elements, in accordance with the principle of order.

In the definition of order-type given by Cantor (see § 157), according to which the order-type of an aggregate is obtained by making abstraction of the particular nature of the elements of the aggregate, it is assumed that the aggregate is given as an ordered aggregate. Again, in his definition of cardinal number (see § 145), Cantor has assumed that the aggregate is given as an ordered one; the cardinal number there appears as the result of a double abstraction, viz. of the particular nature of the elements, and of the order in which they are given. The question however arises, whether the definition of cardinal number should not be such as to be also applicable in the case of aggregates which are not given as ordered aggregates. Cantor has himself, in fact, in his theory of exponentials involving transfinite cardinal numbers, contemplated certain aggregates as having cardinal numbers, whilst such aggregates were not given as ordered aggregates, and *prima facie*, at all events, are not capable of being ordered.

190. Taking the case of an aggregate defined as an ordered aggregate, we now approach the consideration of the fundamental question, whether, and under what conditions, if any, such an aggregate can be regarded as having a definite order-type, and a definite cardinal number. This is equivalent to asking whether, or when, meanings can be given to those terms, in accordance with general definitions, of such a character that they can be treated as permanent objects for thought, or as mathematical entities which may themselves be elements in aggregates.

With reference to Cantor's definition (see § 145), of the cardinal number of a transfinite aggregate, by abstraction, in accordance with which the cardinal number is represented by replacing each element by an abstract unity, it must be observed that such a substitution would replace the given aggregate by another one which had no longer any intelligible relation with the norm by which the original aggregate is defined. The abstract unities would be indistinguishable from one another, and the new aggregate would be indistinguishable from any other non-finite aggregate of such *unities*. It would be impossible to decide, as regards any particular abstract unity, whether it belonged to the aggregate or not; in fact, to make complete abstraction of the individual nature of the elements of an aggregate is to destroy the aggregate. A definition by abstraction could be justified only by the interpretation that abstraction is made of those characteristics only, in which the elements of the aggregate differ from the corresponding elements of all possible equivalent aggregates. Thus the existence of aggregates equivalent to the given aggregate would appear to be essential, if the latter is to be regarded as having a cardinal number to which any definite meaning can be attached. On the grounds stated, the definition of a cardinal number, as the characteristic of a class of equivalent aggregates, is to be preferred to the definition given by Cantor. Accordingly, an aggregate has a cardinal number, only when it is one of a plurality of equivalent aggregates, distinct from one another. In all cases the correspondence between equivalent aggregates must be definable by some norm.

We are thus led to the following statement, containing a definition of cardinal number:

The members of any particular class of equivalent aggregates have a quality in common in virtue of their equivalence. This quality is the cardinal number, and may be regarded as characteristic of each aggregate of the particular class.

A cardinal number has been defined* by B. Russell to be a class of equivalent aggregates. It may then be urged that such a class may contain only one member, and that this is sufficient for the existence of the cardinal number of that member. In fact, Russell infers† the existence of the number $n + 1$ from that of the numbers 0, 1, 2, 3, ... n. Russell objects‡ to the conception of a number as the common characteristic of a family of equivalent aggregates, on the ground that there is no reason to think that such a single entity exists, with which the aggregates have a special relation; but that there may be many such entities, and that there are in fact an infinite number of them. The mind does, however, in point of fact, in the case of finite aggregates at least, recognize the existence of such a single entity, viz. the number, or degree of plurality, of the aggregate; and this creation, by abstraction, is a valid process, subject to the law of contradiction.

191. In Cantor's definition of the order-type of a simply ordered transfinite aggregate (see § 157), abstraction is made of the nature of the elements, their order in the aggregate being alone retained. The order-type is then regarded* as represented by an aggregate of abstract unities, in the order of the elements of the given aggregate. In any ordered aggregate, it is however the individual characteristics of any two elements which determine their relative order in the aggregate, in accordance with some principle of order, valid for the whole aggregate. If complete abstraction be made of the characteristics of the various elements, order has then disappeared from the aggregate. It must be supposed that, in Cantor's representation of the order-type, there are attached to the abstract unities marks of some kind, which may in particular cases be marks indicating position in space or time, by which the order of the various abstract *unities* is denoted, the given aggregate is then really replaced by an aggregate of these marks, and the abstract *unities* are superfluous. These marks, by which order is determined, must also have been associated with the elements of the original aggregate. It thus appears that, in a definition by abstraction, it can be only those characteristics (if any) of the various elements which are irrelevant in determining the order, of which abstraction is made. thus the aggregate is really replaced by a similar one. On these grounds, that definition of an order-type is to be preferred, in which the order-type is defined as the characteristic, or class-name, of a class of similar aggregates. Accordingly, in order that a given aggregate may have an order-type, to which a definite meaning can be attached, it is necessary that the aggregate be one of a plurality of similar aggregates.

---

* *Principles of Mathematics*, vol. I (1903), pp. 111–116. † *Ibid.* p. 497. ‡ *Ibid.* p. 114.

We may accordingly state that

The members of any particular class of similar aggregates have a quality in common, in virtue of their relation of similarity. This quality of mutual similarity possessed by the aggregates is their order-type, and may be represented by a name or symbol, regarded as characteristic of each aggregate of the particular class.

The considerations above adduced may be applied in the case of an aggregate which is a segment of the hypothetical aggregate of all ordinal numbers. In this case it is impossible to make abstraction of the nature of the individual elements of the aggregate, without destroying the order, because the elements are themselves nothing more than marks indicating order.

192. The question has frequently been discussed whether the conception of an ordinal number is necessarily prior to that of a cardinal number. The view is held that the very earliest conception of a number was developed as the result of immediate intuition of the degree of plurality in a very small group of objects, for example of a pair, or of a triplet, or of the unity in a single object, without recourse to the process of counting. In this manner the conception of the first few integral numbers may have arisen, and the conception would be that of cardinal number; the objects in a group not being regarded as in any particular order. This mode of apprehension of the number of objects in a group must certainly have been extremely limited in its scope, and the conception of numbers, after the first few, must have arisen, as explained in § 1, in connection with the process of counting, in which the objects are taken in some definite order; thus leading in the first instance to the ordinal number of the group. It appears certain, then, that the actual order of development of Arithmetic, except for the case of the first few numbers, must have been one in which the ordinal number has a certain priority to the cardinal number.

---

* See *Math Annalen,* vol. XLVI (1895), p. 497.

In the systematic development of the theory of numbers, finite and transfinite, the notion of an ordered aggregate has certainly been fundamental, and not derivative. A reversal of the historical procedure, in which ordered aggregates, having definite ordinal numbers or order-types, were first considered, and afterwards the notion of the cardinal number has arisen, would lead to probably insuperable difficulties. It is no doubt possible to define the cardinal number of a transfinite aggregate, as has in fact been done above, in such a way that the notion of order does not explicitly appear, but the question of the comparability of cardinal numbers so defined at once arises, and the consideration of ordered aggregates would appear to be indispensable for the development of any systematic arithmetic of cardinal numbers. Cantor's Aleph-numbers, as forming an ordered family of cardinal numbers, are essentially dependent upon the theory of the order-types of normally ordered aggregates. It is difficult to see, for example, how it could be known, in respect of a particular aggregate, that it has the cardinal number $\aleph_0$, unless a correspondence of the elements of the aggregate with the ordered set of finite numbers could be shewn to exist. It may be asserted that, since any non-finite part of an enumerable aggregate is itself enumerable, there may exist a part of the integer sequence $1, 2, 3, \ldots n, \ldots$ which cannot be defined by means of a finite set of rules, and yet that such an aggregate must have the cardinal number $\aleph_0$. A distinction is accordingly drawn, by some writers*, between those enumerable aggregates for which a correspondence with the integer sequence $1, 2, 3, \ldots$ can be effectively defined, and those for which this is not the case. It can, however, be shewn that there exists no part of the integer sequence $\{n\}$ which is inherently incapable of finite definition. It has been shewn, in § 63, in the case of the numbers of the continuum, that those numbers, in any interval, that are capable of finite definition, form an aggregate which has all the properties of the continuum itself; and in view of the theory of the order-type of the continuum given in § 163, that aggregate is identical with the continuum.

---

* See, for example, Borel, *Annales sc. de l'école normale*, (3), vol. xxv (1908), p. 447.

Any non-finite part of the integer sequence $\{n\}$ can, however, be made to correspond with a particular number of the continuum in the interval (0, 1); we take such a number expressed as a radix fraction in the binary scale, defined by the rule that the mth digit is 0 or 1, according as the integer m occurs, or does not occur, in the part of the integer sequence which is under consideration. Since there is a (1, 1) correspondence between the parts of the integer sequence and the numbers of the continuum within the interval (0, 1), it follows that, since, as has been shewn in § 63, no number of the continuum is inherently incapable of finite definition, the same is true of the parts of the integer sequence. It thus appears that no enumerable aggregate is inherently incapable of being placed into correspondence with the numbers of the integer sequence, by a finite set of rules, although a particular *fixed* apparatus of definition may be insufficient for this purpose.

193. We proceed to consider those aggregates which consist of ordinal numbers in their order of generation.

There are two distinct methods of establishing the existence of a class of mathematical entities.

(1). Their existence, as definite objects for thought, may be shewn to follow as a logical consequence of the existence of other entities already recognized as existent, or of principles already recognized as valid; so that the existence of the new entities in question cannot be denied without coming into contradiction with truths already known, or at all events with postulations already made. This method may be termed the *genetic method*.

(2). The existence of the entities may be directly postulated; and their mutual relations, and their relations with other entities already assumed to exist, may be defined by means of a complete system of definitions and postulations. Accordingly, the objects in question are a relatively free creation of our mental activity. The validity of the scheme thus set up is established when it is shewn to be free from internal contradiction. Its utility is to be judged by its applicability to the general purposes of the science, and by the light it may throw upon the fundamental principles of that science, in virtue of the scheme containing a generalization of what was previously known. This method may be termed the *method of postulation*. It may, however, be urged that the failure to discover contradictions within a scheme of postulations is no proof that such contradictions do not exist, and that such proof can only be supplied by the exhibition of a system of entities already assumed to exist, such that the relations between them are in accordance with those postulated in the scheme in question.

Both these methods have been employed by Cantor in his theory of transfinite numbers and order-types. In his earlier treatment of the subject, he employed the second of the above methods. The existence of the new number ω, and of the limiting numbers of the second class, was postulated, in accordance with the second principle of generation. Freedom from contradiction, and utility in connection with the theory of sets of points, which suggested the postulations, were relied upon as the grounds upon which the system of new numbers was to be justified. The first number Ω, of the third class, was introduced by a new postulation.

In his later and more abstract treatment of the subject, an account of which has been given in the present chapter, Cantor applied the genetic method. The existence of the number ω is not directly postulated, but is taken to follow from the existence of the aggregate $\{n\}$, of integral numbers; ω is defined to be the order-type of this aggregate, and it is assumed that such order-type is a definite object which can itself be an element of an aggregate. The successive ordinal numbers of the successive classes are obtained by assuming as a general principle, that an ordered aggregate necessarily possesses a definite order-type which can be regarded as itself an object, the ordinal number coming immediately after all those that are the elements of the aggregate of which it is the order-type.

It will be seen later that the assumptions that an ordered aggregate necessarily possesses a definite order-type and that it also possesses a definite cardinal number, both of which can be regarded as objects, lead to the contradiction pointed out by Burali-Forti. It appears, therefore, that the class of entities, which is constituted by the ordinal numbers of all classes, and the similar aggregate of Aleph-numbers, do not satisfy the condition of being subject to a scheme of relations which is free from contradiction. In fact, the principle, in accordance with which their existence is inferred, conflicts with the definition of the aggregates as containing respectively every ordinal number and every Aleph-number. It would then appear that the genetic process which led to the definition of the aggregates of all ordinal numbers, and of all Aleph-numbers, cannot be a valid one without some restriction. Thus the principle that every ordered aggregate has a definite order-type, which may be regarded as a permanent object of thought, cannot be accepted as a universal principle to be used in a genetic mode of establishment of the existence of a class of entities. A denial of the validity of this principle does not however preclude the less ambitious procedure of postulating the existence of definite ordinal numbers of a limited number of classes, in accordance with Cantor's earlier method. So long as the postulation of the existence of ordinal numbers does not go beyond some definite point no contradiction will arise, and the validity of the scheme, for purposes of representation, will suffice to justify the postulations which are made. An attempt to examine the structure of such a class of ordinal numbers, as that of the ωth class, with cardinal number $\aleph_\omega$, or that of the Ωth class, with cardinal number $\aleph_\Omega$, will lead to the conviction that such conceptions are unlikely to prove capable of useful application in any branch of Analysis or of Geometry.

THE PARADOXES OF BURALI-FORTI AND RUSSELL

194. In accordance with Cantor's general theory of ordinal numbers, and of Aleph-numbers, there exist two aggregates

$$1, 2, 3, \ldots\ n, \ldots\ \omega, \omega + 1, \ldots\ \Omega, \Omega + 1, \ldots\ \beta, \ldots$$

$$\aleph_0, \aleph_1, \ldots \aleph_n, \ldots \aleph_\omega, \aleph_{\omega+1}, \ldots \aleph_\Omega, \aleph_{\Omega+1}, \ldots \aleph_\beta, \ldots;$$

the first, the aggregate of all ordinal numbers, which may be denoted by W; and the second, that of all $\aleph$ cardinal numbers. These aggregates are similar to one another, and they contain, respectively, every ordinal number, and every cardinal number which belongs to a normally ordered aggregate.

The aggregate W is normally ordered. For, if W_1 be any part of W, such that W contains one or more elements of higher order than all the elements of W_1, it follows that W_1 is a part of some segment of W that also contains one or more such elements. Since such segment is normally ordered, there exists one element of that segment, and therefore of W, which immediately follows the part-aggregate W_1. It is thus seen that, since the conditions of § 165 are satisfied, W is normally ordered. In accordance with the principle which is fundamental in the whole theory, that every normally ordered aggregate has a definite order-type, which is its ordinal number, and has also a definite cardinal number, the aggregate W has an ordinal number γ, and a cardinal number $\aleph_\gamma$. The ordinal number γ must itself occur in the aggregate W, and must therefore be the greatest ordinal number, *i.e.* the last element of W. There can, however, be no last ordinal number; for, on the assumption of the existence of γ, an aggregate, of ordinal number $\gamma + 1$, can be formed, by adding to W a new element e, of higher rank than all the elements of W. Therefore a contradiction has been arrived at.

This contradiction in the conception of the aggregate W, of all order-types of normally ordered aggregates, was pointed out by Burali-Forti*, but as is stated by Jourdain†, it was discovered by Cantor in 1895, and communicated by him to Hilbert in 1896, and to Dedekind in 1899. The contradiction has been discussed by various writers, and the conclusions they draw from its existence are not in full agreement with one another. Burali-Forti inferred from it that ordinal numbers are not comparable with one another, but this is refuted by Cantor's theorem in § 168. Jourdain shews that W is normally ordered, but concludes that it possesses neither order-type nor cardinal number. He describes W as an "inconsistent" aggregate, a term which had been previously employed by Cantor in the same connection. It would thus appear that there are aggregates which have no order-type, and no cardinal number, and that the above aggregates belong to such class. Such aggregates are called inconsistent aggregates, in virtue of the fact that the act of synthesis, by which they are formed, cannot be carried out without leading to a conception which contains an element of contradiction.

---

* *Rend. del circolo mat. di Palermo*, vol. XI (1897), p. 164.

† *Phil. Mag.* (6), vol. VII (1904), pp. 67, 70.

It was pointed out by Schoenflies* and by Bernstein†, that the hypothetical aggregate W lacks one property that belongs to all its segments, and to all normally ordered aggregates, viz. that a new normally ordered aggregate can be formed by the addition of a new element, of higher rank than all those of the original aggregate. In fact the first principle of generation of the successive ordinal numbers in Cantor's scheme is that to each ordinal number there follows a next one, and it is clearly impossible that the aggregate of *all* ordinal numbers can have an ordinal number that is subject to this principle. It thus appears that W can have no ordinal number, or order-type. This being so, the question, whether W can be properly said to be an aggregate, or not, depends on the precise meaning assigned to the term aggregate. It would appear desirable that such a limitation should be given to the definition of the term that such hypothetical aggregates as W would not fall under the definition. All ordinal numbers form a class, and thus doubt is thrown upon the advisability of making the term aggregate synonymous with that of the "class" of Logic. Moreover, it would appear desirable that the definition of an aggregate should be so restricted that every aggregate has a cardinal number and, if normally ordered, an order-type. It has been pointed out‡ by Mirimanoff that the antinomy of Burali-Forti would arise, even if the existence of all the ordinal numbers in the scheme of Cantor be not admitted. Instead of W, the hypothetical aggregate of all existing ordinal numbers could be contemplated, and the same reasoning would shew that it would be contradictory to ascribe an ordinal number, or order-type, to this aggregate. Let us assume, for example, that the finite ordinal numbers were the only ordinal numbers whose existence was admitted; it would then be contradictory to regard the totality of all these ordinal numbers as having an ordinal number, or order-type; in fact this totality would be an "inconsistent" aggregate, in the sense employed by Cantor and Jourdain.

The mode of generalizing the ordinal numbers in Cantor's scheme requires a fresh postulation of existence, in the case of the lowest numbers of each of the successive classes. Although the contradiction of Burali-Forti cannot be invoked to shew that such a postulation leads to contradiction, as long as we refrain from considering such an aggregate as that of *all* ordinal numbers, or of *all* existing ordinal numbers, it cannot be directly shewn that at no stage of the successive postulations does some contradiction arise. It would appear then that the validity of the whole scheme, being subject to the possibility of the discovery of contradictions at some stage or other, the existence of the numbers in Cantor's scheme must be regarded, from the point of view of the Mathematician, only as a working hypothesis. The utility of such hypothesis must be judged by its consequences and its applications in Analysis.

---

* *Bericht*, 2ter Theil, p. 27.

† Bernstein, *Math. Annalen*, vol. LX (1905), p. 187.

‡ See *L'enseignement mat.*, Jan. to March, 1917.

195. The antinomy* of Russell depends upon the consideration that two species of aggregates may be contemplated. An aggregate M, of the first species, does not contain M itself as an element, but an aggregate M of the second species does contain M itself, as an element. An example of an aggregate of the second species is the aggregate of all aggregates. It is clear that an aggregate of the second species contains an element which is itself an aggregate of the second species It follows that an aggregate, all of whose elements are aggregates of the first species, is itself of the first species. The aggregate which contains, as elements, all aggregates of the first species is a notion essentially affected with contradiction. For it must be itself of the first species, since all its elements are of that species; on the other hand, it must be of the second species, since it must contain itself as element, for the aggregate itself is one of the aggregates of the first species.

It thus appears that the aggregate of all aggregates of the first species cannot exist as a notion free from contradiction.

This contradiction may be considered in relation both to Cantor's definition of an aggregate, and to the definition of an aggregate as a class. In accordance with Cantor's definition, the elements of an aggregate, being logically prior to the aggregate, must be definable in a manner which does not make use of the aggregate itself in the definitions; and thus it would appear that no aggregate which contained itself as an element would fall under Cantor's definition of an aggregate. On the other hand, the definition of an aggregate as a class would not *prima facie* preclude the contemplation of such a conception as the class of all classes.

It is clear that the proper definition of an aggregate should be such as to exclude all hypothetical aggregates which contain elements that cannot be defined without employing the aggregate itself; in the language of Poincaré, non-predicative definitions should be excluded.

196. With a view to the elucidation and generalization of the antinomies of Burali-Forti and Russell, a classification of aggregates has been given* by Mirimanoff, according to the modes in which their elements are composed of other elements. The elements m, of an aggregate M, may be simple elements which are not decomposable into other elements. On the other hand, an element m may itself be an aggregate composed of elements e; thus $m \equiv \{e\}$. Again, the elements e, of which m is composed, may themselves be aggregates composed of other elements e'; thus $e \equiv \{e'\}$.

---

* See *Principles of Mathematics*, vol. I (1903), chap. X.

It is clear that this process of decomposition of the elements of an aggregate into aggregates of other elements, and of these elements into aggregates of other aggregates, may proceed until only simple elements, no longer decomposable, are obtained; or on the other hand the process may go on indefinitely.

An aggregate is said, by Mirimanoff, to be an *ordinary* aggregate, when each element m is such that, after a finite number of decompositions, simple undecomposable elements are obtained. An aggregate which does not possess this property is said to be an extraordinary aggregate.

Two aggregates are said to be isomorphic when (1), they are equivalent, *i e.* when there is a (1, 1) correspondence between their elements, and (2), when, m, m' being two corresponding elements, which are decomposable, there is also equivalence between the two aggregates of which they consist, and when such equivalence also holds in case further decomposition can take place, and so on, so long as any continued decomposition of elements can be made.

An aggregate is said to be of the first species if it is not isomorphic with any of its elements, and it is said to be of the second species if it is isomorphic with one at least of its elements. The hypothetical aggregate, which consists of all aggregates of the first species, or also that of all ordinary aggregates, involves the same contradiction as in the case of Russell's antinomy.

An aggregate of the second species is an extraordinary aggregate, but an aggregate of the first species may also be extraordinary.

THE MULTIPLICATIVE AXIOM

197. An axiom, known variously as the "multiplicative axiom," the "principle of Zermelo," and the "general principle of selection," which can be expressed in several equivalent forms, was first stated explicitly in connection with a proof given by Zermelo†, in 1904, that every aggregate can be ordered in normal order; although it had been implied in the reasoning of various writers in connection with special theorems in the theory of sets of points, and in the theory of functions. The axiom has been stated‡ by Zermelo in the following form:

---

* *L'enseignement mat*, Jan. to March, 1917, p. 42.

† *Math. Annalen*, vol. LIX (1904), p. 514.

‡ *Math. Annalen*, vol. LXV (1908), p. 110.

An aggregate S, which falls into an aggregate of separate parts $A, B, C, \ldots$, each of which contains at least one element, possesses at least one part S, which has in common with each of the parts $A, B, C, \ldots$, exactly one element.

In accordance with the definition, given in § 149, of the product of a finite, or infinite, set of aggregates, as an aggregate of which each element consists of an association of elements, one from each of the aggregates of the given set, the above principle is equivalent to the assertion that the product of the aggregates of the given set contains at least one element; and thus the principle may be termed the "multiplicative axiom," in that it asserts the existence of the product.

The principle can be stated in the less objective form that a choice can be made of a single element, from each one of a set $\{M\}$ of aggregates M; an aggregate thus being formed by an infinite number of such acts of choice. When viewed in this manner the principle is spoken of as the "general principle of selection" (das Prinzip der Auswahl).

In the discussions to which this principle, or axiom, has given rise, much divergence of opinion on the part of mathematicians has emerged. Borel, who pointed out that Zermelo's proof that an aggregate can be normally ordered simply establishes that this theorem is equivalent to the principle of selection, distinguished* between an enumerable and an unenumerable set of acts of choice, and rejected, as outside the domain of Mathematics, all reasoning founded upon the supposition of an unenumerable set of acts of choice. The impossibility of proving the principle was expressly stated by Zermelo, and emphasized by Borel, and by Peano†. Poincaré, in the course of his discussion‡ of the "Logistic" of Peano and Russell, expressed the view that the principle, although incapable of proof, is an indispensable axiom.

198. By some Mathematicians, the employment of an infinite set of acts of choice, even if that set be enumerable, is rejected as outside the domain of Mathematics, on the ground that the hypothetical entity, the existence of which is asserted in such a case by the principle, is not properly defined.

In the course of a discussion § by Hadamard, Borel, Baire, and Lebesgue, the distinction was taken into account between independent acts of choice and such as are dependent upon acts of choice that have been previously made.

In those special cases in which the aggregates are of such a nature that it is possible effectively to define an aggregate which consists of one element belonging to each of the given aggregates, the principle is not required as an axiom, but such special cases may be regarded as instances in which the truth of the principle is verified. In a case in which we are unable to give such effective definition, the principle amounts to an assertion of the existence of the aggregate in question, apart from any question of effective definition, and at least asserts that we are warranted to make the same deductions as if we were in possession of such effective definition. In this connection it may be pointed out that there is a certain relativity in the notion of effective definition, as has already been shewn (§ 63) in the case of the definition of numbers of the arithmetic continuum; in fact, the possibility of effective definition of an object is relative to an apparatus of definition which cannot be considered as once for all, and finally, fixed.

---

* See *Math. Annalen*, vol. LX (1905), p. 195.

† *Rivista di Mat.* vol. VIII, No. 5, p. 145.

‡ See the *Revue de Métaphysique et de Morale*, vols. XIII (1905), p. 816 and XIV (1906), p. 866.

§ "Cinq lettres sur la théorie des ensembles," *Bull. de la soc. math. de France*, vol. XXXIII (1905), p. 261.

Even if the number of aggregates from which the selection is to be made be finite, there remains a question as to the scope of the axiom. It is sufficient to consider whether, and in what sense, it is always possible to select a single element from a single given aggregate. Whether this can be done effectively will depend upon the mode in which the aggregate is defined. For the possibility of such choice, it must in the first place be known that the aggregate is not a null-aggregate, *i.e.* that it contains at least one element. The assumption of this knowledge being made, the effective determination of an element may be impracticable, or difficult; but this will not affect its logical possibility. It appears, however, that there are cases in which even the logical possibility of determination cannot be demonstrated; and in such cases recourse must be had to the axiom of existence, if an element of the aggregate in question is required for the purpose of any process of reasoning in which such element is employed.

Some of the objections raised to the axiom amount to the assertion that the axiom is meaningless, if it is interpreted as asserting the existence of an aggregate of which the elements cannot, even in theory, be effectively defined. The difference of opinion upon the point, as to whether, or not, an aggregate can be said to exist, when none of its elements can be effectively defined, would appear to reveal a fundamental difference of view on a matter of Ontology, which cannot be resolved within the domain of Mathematics, as it involves an irreconcilable divergence of attitude in general Philosophy. It was pointed out by Hadamard (*loc. cit.*) that the real question at issue is whether it is possible to demonstrate the existence of mathematical entities which cannot be precisely defined; a question which Hadamard himself answers in the affirmative, though with this answer the other writers do not appear to be in agreement. By some mathematicians, the principle is accepted in a pragmatic sense, as a useful instrument of Mathematical research, and as a guide which suggests results that may be verifiable by other means, although the principle cannot be regarded as standing upon an absolutely firm basis.

The case for the acceptance of the axiom rests partly upon the fact that the postulation of its validity had been implicitly made, as self-evident, by many writers, in various investigations in the theory of sets of points, and in the theory of functions, before it was stated explicitly; and that it has since not only proved fruitful in many further developments of those theories, but that results obtained by its use have been verified in many important cases. A detailed analysis has been given* by Sierpiński of general results in the theories of sets of points, and of functions, which cannot be established without the aid of the axiom, and of other cases in which the employment of the axiom is not necessary. The method of G. H. Hardy (§ 186), for the construction of a set of points of cardinal number $\aleph_1$, is a case in which the principle of selection is required, since there exists an indefinite number of sequences of the ordinal numbers, preceding a limiting ordinal number γ, of each of which γ is the limiting number. It is not possible to give a norm by which these sequences can be determined.

199. A limiting point P, of a set of points G, has, in § 52, been defined as a point such that in every neighbourhood of P there are points of G. It has been pointed out by Sierpiński that, in the general case, the inference, from this definition, that there exists a sequence $P_1, P_2, \dots P_n, \dots$ of points of G, which converges to P as sole limiting point, so that the distance PP_n converges to zero, as n is indefinitely increased, cannot be made without having recourse to that case of Zermelo's axiom in which the selections are made from an enumerable sequence of sets of points.

For simplicity, let us consider the case in which G is a linear set; this differs in no essential respect from the case in which G is a set in p dimensions, where $p > 1$. Let us suppose that the point $x = 0$ is a limiting point of G, in accordance with the definition of § 52. Let G_n be that part of G which consists of points x that satisfy the conditions $\frac{1}{n+1} \leqq |x| < \frac{1}{n}$. Thus G_n is in the two half-closed intervals $\left(\frac{1}{n+1}, \frac{1}{n}\right), \left(-\frac{1}{n}, -\frac{1}{n+1}\right)$. Consider now the sequence of sets $G_1, G_2, \dots G_n, \dots$ of which the sum is the relevant part of G. The axiom of Zermelo asserts the existence of a set of points $P_r, P_{r+1}, \dots P_n, \dots$; such that P_n belongs to G_n, for all the values $r, r+1, \dots$ of n. Unless it is possible to define by a norm such a set of points, its existence can only be assumed in virtue of the axiom. This set of points is a sequence such as is required, which converges to the point $x = 0$.

---

* *Bull. de l'acad. des sciences de Cracovie*, April—May, 1918.

There is one important case in which Zermelo's axiom is not required, viz. when G contains a known enumerable set H, for example the set of rational points, that is everywhere dense in G. The point P_n may then be defined as that point of H, of lowest rank when H is arranged in enumerable order of type ω, which belongs to G_n.

It has been shewn conversely by Sierpiński that the following theorem follows as a consequence of the assumption that, if P be a limiting point of a set G, in accordance with the definition in § 52, there exists a sequence $\{P_n\}$, of points of G, which converges to P as its sole limiting point.

If $G_1, G_2, \ldots G_n, \ldots$ *be sets of points, no two of which have an element in common, there exists a sequence* $P_1, P_2, \ldots P_n, \ldots$ *of points, such that each one belongs to one of the sets of* $\{G_n\}$ *and that no two of them belong to the same set of* $\{G_n\}$.

By means, for example, of the transformation

$$\xi = \frac{1}{2n(n+1)}\left(\frac{x}{1+|x|} + 2n + 1\right),$$

the set G_n can be put into correspondence with a set Q_n, all the points ξ of which are interior to the interval $\left(\frac{1}{n}, \frac{1}{n+1}\right)$. Let Q denote the set of points ξ, given as the sum of all the sets $Q_1, Q_2, \ldots Q_n, \ldots$; then the point $\xi = 0$ is a limiting point of the set Q.

Now let it be assumed that there exists a sequence $\{q_m\}$, of points of Q, which converges to $\xi = 0$ as its sole limiting point. We may suppose that there is only one of these points in any one set Q_n, for, if q_m belongs to a set of $\{Q_n\}$ to which any one of the points $q_1, q_2, \ldots q_{m-1}$ belongs, we may remove q_m from the sequence of points. Let us suppose that q_n belongs to $Q_{\bar{n}}$; then by applying the inverse transformation to that of x into ξ, we have a sequence of points $P_1, P_2, \ldots P_n, \ldots$ which belong to the sets

$$G_1, G_2, \ldots G_{\bar{n}}, \ldots$$

respectively.

200. The question of the justification of the assumption that every infinite aggregate contains an enumerable aggregate as a part is related to the principle of selection. It was assumed by Cantor that it is always possible to pick out of a given aggregate, successively, elements that correspond to the numbers of the integer sequence. In a large class of cases this can be effected by means of a suitable norm, but in the general case an infinite number of acts of choice, each of which is affected by the choices already made, are required. In case, however, a given aggregate M can be divided into an enumerable set of parts $M_1, M_2, M_3, \ldots$, no two of which have an element in common, one element may be selected from each of these parts, in accordance with the principle of selection; and thus the existence of the enumerable part of M follows as a consequence of the principle. In case M is an aggregate of points, in one or more dimensions, an enumerable set of non-overlapping cells, or intervals, may be defined, each of which contains a part of M; and we may take the parts

$$M_1, M_2, M_3, \ldots,$$

of M, to be those parts that are contained in the cells, or intervals, of the set.

MULTIPLE CORRESPONDENCE

201. *À propos* of a criticism[*], by Levi, of Bernstein's proof (see § 185), that the aggregate of all closed sets in the p-dimensional continuum has the power c of the continuum, Bernstein[†] has endeavoured to avoid the difficulty involved in the use of a correspondence which cannot be effectively defined, by introducing the conception of "multiple equivalence." Thus, if there are two aggregates M and N, for which an aggregate $\Phi = \{\phi\}$ of reversible (1, 1) correspondences ϕ exists, in which no element is special (ausgezeichnet), then the two aggregates are said to be multiply equivalent. The cardinal number $\overline{\overline{\Phi}} \equiv f$ is then termed the multiplicity of the correspondence. In case the multiplicity is unity, the aggregates are said to have a one-valued correspondence. The difficulty of this conception is the same as that in the multiplicative axiom itself, viz., that the aggregate Φ employed is such that no single element of it is capable of definition, and that the elements are consequently indistinguishable from one another.

THE NORMAL ORDERING OF AN AGGREGATE

202. It was regarded by Cantor as highly probable that every aggregate is capable of being normally ordered; and this is equivalent to the theorem that every transfinite cardinal number is an Aleph-number. If this principle could be accepted as valid, the particular theorem would follow, that the arithmetic continuum is capable of being normally ordered; and the only question which would remain open, as regards this aggregate, would be as to which particular Aleph-number is the cardinal number c of the continuum.

No proof of the correctness of his surmise was published by Cantor himself, but a proof was given by Jourdain[‡], which depends upon an argument in which the "inconsistent" aggregate W (§ 194) was employed to shew that no cardinal number can be greater than every Aleph-number. Apart from the employment of the concept of the aggregate W, which has been shewn to be affected with contradiction, this proof involved implicitly the principle of selection.

---

[*] *Lomb. Ist. Rend.* (2), 1902, p. 863.

[†] *Göttinger Nachrichten*, 1904, p. 557.

[‡] *Phil. Mag.* (6), vol. VII (1904), p. 67; see also *Math. Annalen*, vol. LX (1905), p. 465.

A second, and more cogent, proof was given* by Zermelo, in connection with which the axiom associated with his name was explicitly stated as necessary for the purpose of his proof. He has later published† a new form of the proof, together with a reply to criticisms of his earlier proof.

It is assumed that, in each part M', of a given aggregate M, one element m', called the special (ausgezeichnetes) element of M', can be chosen. The possibility of doing this for all the parts of M follows from the axiom. Each element M', of $\{M'\}$, corresponds to a special element m' which belongs to M; and this particular mode of distributing the elements of M upon the elements of $\{M'\}$ is called a "covering" γ; the employment of a particular "covering" γ is essential to the proof. A γ-aggregate is then defined as follows:—Let M_γ be a normally ordered aggregate consisting of different elements of M, such that, if a be any arbitrarily chosen element of M_γ, and if A be the segment of M_γ defined by a, which segment consists of all the elements of M_γ that precede a, then a is always the special element of $M - A$. Every such aggregate M_γ is a γ-aggregate. If every element of M which occurs in a γ-aggregate be called a γ-element of M, it is shewn that the aggregate L_γ, of all γ-elements, can be so ordered that it is itself a γ-aggregate, and contains all the elements of the original aggregate M. It follows that M can be normally ordered.

It will be observed that the theorem that any aggregate can be normally ordered does not assert the possibility of effectively carrying out, in the case of any given aggregate, the process of arranging the elements of the aggregate in normal order. The theorem must rather be regarded as an existence theorem, deducible from certain postulates which include the multiplicative axiom.

In the later form of his proof, Zermelo states the theorem in the following manner:

If there corresponds to each part of the aggregate M, one element of that part, its special element, then the aggregate $U(M)$, which contains as its elements all parts of M, has one and only one element $\bar{M}$, such that there corresponds to each part P, of M, one and only one element P_0, of $\bar{M}$, such that P_0 has P as a part, and an element of P as its special element. The aggregate M is normally ordered by means of $\bar{M}$.

The existence of the aggregate $U(M)$ as a valid conception is assumed.

THE COMPARABILITY OF AGGREGATES

203. From the theorem that every aggregate can be normally ordered, it would follow that the cardinal number of any infinite aggregate must be an Aleph-number, or the equivalent theorem of the "comparability of aggregates," that of any two aggregates one at least is equivalent to a part of the other. It has been proved* by Hartog, conversely, that if the comparability of every pair of aggregates be assumed, the theorem that any aggregate can be normally ordered can be deduced. This theorem is established on the basis of a system of axioms given† by Zermelo.

---

* *Math. Annalen*, vol. LIX (1904), p. 514. † *Math. Annalen*, vol. LXV (1908), p. 107.

It follows from Hartog's result, combined with Zermelo's theorem, that the three principles of selection, of comparability of two aggregates, and of the normal ordering of an aggregate, are completely equivalent to one another, in that any two of them can be deduced from the third. In case none of the three principles is assumed to be valid, Hartog's investigation proves that there exists no aggregate whose cardinal number is greater than all the Aleph-numbers. The proof depends upon a classification of all the normally ordered aggregates of which the elements are also elements of a given aggregate M; the validity of the conception of the aggregate of all such normally ordered aggregates, corresponding to a given aggregate M, being assumed.

204. The discussion of the fundamentals of the theory of aggregates, given in the latter part of the present chapter, in which the divergence of view amongst mathematicians, as regards some parts of the theory, has been indicated, makes it clear that, for some time to come, a critical attitude is likely to be maintained in some quarters as regards the theory of the transfinite, at least in some of its developments. The extreme sceptical attitude has been expressed by Poincaré, in the words‡ "Il n'y a pas d'infini actuel; les Cantoriens l'ont oublié, et ils sont tombés dans la contradiction." Even if the general theory of classes of order-types and of Aleph-numbers be regarded by many mathematicians as still in the region of speculation, nevertheless the debt which Mathematical Science owes to the genius of G. Cantor will not be materially diminished. The fundamental distinction between enumerable and unenumerable aggregates, the interpretation of the arithmetic doctrine of limits, the ordinal theory of the arithmetic continuum, and the conception of the ordinal numbers of the second class, with their application to the theory of sets of points, remain as permanent acquisitions, independently of the acceptance of the whole of the higher developments of the abstract theory of aggregates. This order of ideas has become indispensable, for purposes of exact formulation, in Analysis and in Geometry; it is constantly receiving new applications, owing to its admirable power of providing the language requisite for expressing results in the theory of functions with the highest degree of rigour and generality. Cantor's creations have rendered inestimable service in formulating the limitations to which many results in

---

* *Math. Annalen*, vol. LXXVI (1915), p. 438.

† *Math. Annalen*, vol. LXV (1908), p. 261.

‡ *Revue de Métaph. et de Morale*, vol. XIV (1906), p. 316.

Analysis, formerly supposed to be universally valid, are subject. The outlying parts of the theory, although they open up a most interesting field of investigation, from which much may be hoped for in the future, do not appear as yet to be comparable in importance, for the general purposes of Analysis, with those parts which are accepted as fully established by all except the most extreme finitists. The fact that the general theory of Aleph-numbers, as an abstract development of the theory of order, has received but few applications in the theory of functions, differentiates it from the theory of normally ordered enumerable aggregates, which has now become a most useful instrument of discovery in the theory of functions of one or more variables. All aggregates of points of a continuum, which are defined by the methods in ordinary use, have either the power of the aggregate of rational numbers, or else that of the arithmetic continuum itself. The theories of these two kinds of aggregates, including as they do a complete arithmetic theory of limits, would thus appear to afford a sufficient basis for all the ordinary parts of Analysis.

CHAPTER V

FUNCTIONS OF A REAL VARIABLE

205. If we suppose that an aggregate of real numbers is defined, the aggregate being either enumerable, or of the power of the continuum; such an aggregate is said to be the domain, or field, of a real variable. It is necessary for the purposes of Analysis to be able to make statements applicable to each and every real number of the aggregate, and which shall be valid for any particular number that may, at will, be selected. This is done by employing the *real variable*, denoted by some symbol other than those used to denote real numbers, and the essential nature of the variable consists in its being identifiable with any particular number of its domain. The symbols used for denoting variables differ from those employed in the case of numbers in being non-systematic. Operations involving real variables $x^{(1)}, x^{(2)}, x^{(3)}, \ldots$, with, or without, particular numbers, are carried out in conformity with the same formal laws as hold in the arithmetic of real numbers. The result of any such operation is itself a variable with a domain of its own, which may, or may not, be identical with that of any of the constituent variables.

The numbers being used to designate in the usual manner the points of a set on a straight line, the variable may then be taken to refer to the points of the set.

If the given set of points be bounded, in the sense explained in § 47, then the domain of the variable is said to be *limited*, or *bounded*. When the domain of the variable is not limited, it is said to be *unlimited*, or *unbounded*, in one or in both directions.

The variable is said to be continuous, in a given interval (a, b), when all the points of the interval, including a and b, belong to the domain of the variable. If the points a, b do not belong to the domain, but every internal point of the interval does so belong, the variable is said to be continuous in the *open* interval (a, b), or *within* the interval (a, b). The corresponding definitions apply to the case of an aggregate of any number p of dimensions, which is regarded as the domain of p independent variables $x^{(1)}, x^{(2)}, \ldots x^{(p)}$. In this case a closed, or an open, connex domain of p dimensions takes the place of a closed, or an open, interval. A variable point $(x^{(1)}, x^{(2)}, \ldots x^{(p)})$, of such a domain, is often conveniently represented symbolically by a single variable x.

The term "variable" has been commonly associated with the conception of a point moving in a straight line or in a curve. It has however been pointed out in the course of the discussions of the continuum, contained in the earlier Chapters, that the continuum cannot legitimately be regarded as a synthetic construction formed by a set of points determined successively. Successive determination is applicable only in the case of any enumerable sequence which may be defined within the continuum, and such a sequence may represent a succession of positions of a point moving in a straight line. It is however unnecessary to proceed to a detailed analysis of the conception of motion, because the Theory of Functions has no need of the conception of temporal succession. The theory makes continual use of simply infinite sequences determined in the continuum; and any such sequence may be regarded as a series of distinct determinations of the variable in which the elements are in logical succession, each element after the first being preceded and succeeded by definite elements.

THE FUNCTIONAL RELATION

206. If, to each point of the domain of the independent variable x, there be made to correspond in any manner a definite number, so that all such numbers form a new aggregate which can be regarded as the domain, or field, of a new variable y, this variable y is said to be a (single-valued) *function* of x. The variables x, y are called the independent and the dependent variable respectively; and the functional relation between these variables may be denoted symbolically by the equation $y = f(x)$. In this definition no restriction is made *a priori* as regards the mode in which, corresponding to each value of x, the value of y is assigned; and the conception of the functional relation contains nothing more than the notion of determinate correspondence in its abstract form, free from any implication as to the mode of specification of such correspondence. In any particular case, however, the special functional relation must be assigned by means of a set of prescribed rules or specifications, which may be of any kind that shall suffice for the determination of the value of y corresponding to each value of x. Such rules may, in any particular case, be embodied in a single arithmetic formula, from which the value of y corresponding to each value of x is arithmetically determinable; or the rules may be expressed by a set of arithmetic formulae each one of which applies to a part of the domain of the independent variable. In case these formulae be reducible to a set of mutually independent formulae, that set must be a finite one. In case the function be defined by an enumerably infinite set of formulae, each applicable to a part of the domain, these formulae cannot be mutually independent, but must be subject to some norm.

It should be observed that when, for any particular value of x, the corresponding value of y is given by means of any arithmetic formula, the numerical value of y is in general only formally determinate; for only a finite number of elements of a convergent aggregate which defines the value of y can in general be actually found, and thus the value of y can be specified only to any required degree of approximation, but it is still regarded as perfectly determinate.

The domain of x consisting of a set (P) of points, the values of y, in the case of a given functional relation $y = f(x)$, may be represented by points Q on a linear interval; all such points forming a linear set (Q). The set (Q) is said to be the functional image of the set (P), determined by the function $f(x)$; to each point of (P) there corresponds a single point of (Q), if $f(x)$ be a single-valued function; but to each point of (Q) there may correspond a finite, or an infinite, number of points of (P).

If x be employed to denote a variable point $P(x^{(1)}, x^{(2)}, \dots x^{(p)})$ in a p-dimensional field, all that has been said will apply to the case of a function y, of p independent variables. The linear set (Q) will be the functional image of the p-dimensional set (P).

The perfectly general definition of a function which has been given above is the culmination of a process of evolution which has proceeded largely in connection with the study of the representation of functions by means of trigonometrical series. By the older mathematicians, a function was understood to mean a single formula, at first usually only a power of the variable; but afterwards it was regarded as defined by any one analytical expression, and was extended by Euler to include the case in which the function is given implicitly by a formal relation between the two variables. In connection with the problem of the determination of the forms of vibrating strings, which led to the discussion of functions represented by trigonometrical series, the conception arose of a single function defined in different intervals by means of different analytical expressions. The arbitrary nature of a function given by a graph was distinctly recognized by Fourier; thus the notion of a function was emancipated from the restriction that an *a priori* representation of it is necessary by a single formula.

The idea that a function can be defined completely, in the case when the domain of the independent variable is a finite continuous interval, by means of a graph, arbitrarily drawn, leaves out of account the essentially unarithmetic nature of geometrical intuition. A curve that is drawn is indistinguishable by the perception from a sufficiently great number of discrete points; and thus all that is really given by an arbitrarily drawn graph consists of more or less arithmetically inexact values of the ordinates at those points of the x-axis at which we are able to measure ordinates. In order that a curve may be really known, sufficiently to serve for the purpose of defining a function, a series of rules must be prescribed, by means of which the values of the ordinates can be formally determined at all points of the x-axis. It is sometimes said, in order to illustrate the generality of the functional relation, that a function is definable in the form of a table which specifies values of y corresponding to values of x, this table being of a perfectly arbitrary character. The inadequacy of such illustration is manifest, if we consider that, even if the table were an endless one, as has been remarked in § 187, no aggregate of y-values can be defined by an endless set of numbers, apart from the production of a norm by which those numbers are defined. Moreover, even if the table were subject to a definite norm, it could only theoretically suffice to define a function of a variable whose domain consisted of an enumerable set of points, and would be totally inapplicable to the case in which the variable has a continuous domain, unless some special restrictive assumptions as to the nature of the function be introduced, by means of which the values of the function are made determinate at the remaining points of the continuous domain.

It thus appears that an adequate definition of a function for a continuous interval (a, b) must take the form first given to it by Dirichlet*, viz. that *y is a single-valued function of the variable x, in the continuous interval (a, b), when a definite value of y corresponds to each value of x such that $a \leq x \leq b$, no matter in what form this correspondence is specified.* A particular function is actually defined when y is arithmetically defined for each value of x.

The theory relating to the properties of functions which retain their complete generality, in accordance with the abstract definition given above, has been, at present, but little elaborated, since comparatively few deductions of importance can be made from that definition which will be valid for all functions. When, however, the nature of a function is in some way restricted, either in the whole domain, or in the neighbourhoods of special points of that domain, there is room for the development of a detailed theory which shall deal with the peculiarities that follow from such restriction upon the complete generality of functions.

207. The functions defined in accordance with the above definition are known as *single-valued* functions, since, to each value of x in the domain of x, there corresponds a single value of y. The definition may be so generalized as to be applicable to *multiple-valued* functions. This is done by replacing the requirement that, to each value of x in the domain of x, there shall correspond a single value of y, by the more general statement that, to each value of x there shall correspond a definite aggregate of values of y. The aggregate of values of y may, for any particular value of x, consist of a finite, or of an infinite, set of numbers. A particular function is then defined when the aggregate of values of y is arithmetically determinate for each value of x, in accordance with the criteria for the determinancy of a linear aggregate which have been developed in the theory of aggregates. Although the Theory of Functions, as developed in the present work, is mainly concerned with single-valued functions, it is necessary, or at least convenient, in the course of the examination of particular functions and classes of functions, to make use of auxiliary functions which are multiple-valued at certain points of the domain of the independent variable. Moreover, Dirichlet's definition, in its original form, has the inconvenience that it excludes from the category of functions those represented by analytical expressions which, for particular values of the independent variable, cease to define a single number. For example, an infinite series which, for particular values of the variable, either diverges, or ceases to converge to a single definite limit, does not define a single-valued function in accordance with Dirichlet's definition, for the whole domain of the variable, and yet it is convenient so to extend the meaning of the term "function" that a function may be defined for the whole domain by such a series.

---

* See Dirichlet's *Werke*, vol. I, p. 135.

The distinction has been considered in detail by Brodén* between those functions for which the relation between the dependent variable y and the independent variable x is formally the same for the whole domain of x, and those functions for which the domain of x is divisible into a plurality of parts, for which the forms of the relation between x and y are different. He remarks that the distinction is one relating to the character of the definitions rather than to the nature of the functions themselves; in the former case the function is said to be *homonomically* defined, and, in the latter case, to be *heteronomically* defined. Brodén has given a formal proof that, when a function is heteronomically defined, the number of parts into which the domain of x can be divided, so that the relations of y to x in any one part are completely independent of the relations in the other parts, must be finite.

The Theory of Functions of a Real Variable is concerned with the classification of functions, according as they possess various special properties, *e.g.* continuity, differentiability, integrability, throughout the domain of the independent variable, or at, or near, special points which form part of that domain. The theory requires the introduction of precise arithmetical definitions of the scope and meaning of these characteristic properties, and is largely concerned with the determination of criteria which shall suffice to decide, in the case of a function defined in some special manner, what can be inferred as regards the possession by such function of properties other than those that are immediately apparent from the definition itself. Much of the theory is concerned with a minute examination of functions, and of classes of functions, which possess properties that do not occur in the case of those functions which are employed in ordinary analysis and in its applications to Geometry and Physics; and the theory has in consequence frequently been described as the Pathology of Functions. It appears however, from the theory itself, that many of those peculiarities, which from the point of view of traditional Analysis would be described as exceptional, have no claim to be so described; that in fact it is in the functions of ordinary Analysis that the abnormalities really occur, such functions occupying an exceptional position in relation to a scientific analysis of the properties of functions in general. An important result of the labours of those who have developed the modern theory of functions of a real variable has been that restrictive assumptions, which had previously been unconsciously made in the processes of ordinary Analysis, have been placed in a clear light; and it has been shewn that modes of reasoning which had their origin in an uncritical application of ideas obtained from spatial intuition would fail to yield correct results when applied to cases of unrestricted generality; the unsoundness of the logical basis of such reasoning being thereby demonstrated.

---

* *Acta Univ. Lund.* vol. XXXIII (1897), "Functionentheoretische Bemerkungen und Sätze."

In ordinary Analysis the domain of the independent variable is taken to be a limited, or unlimited, continuous interval. In the theory of functions, on the other hand, it has been found advantageous to consider also the properties of functions defined for a domain which is not a continuous one. It appears, in particular, that a non-dense perfect set of points, or more generally any closed set, is well suited to be the domain of a function, inasmuch as, for such domains, the principal peculiarities of functions, such as continuity, differentiability, &c., are capable of precise formulation, and can serve for purposes of classification, exactly as in the case of functions defined for a continuous domain. Much of the recent progress in the subject is due to a recognition of the parity of all perfect sets of points, not only as regards their internal structure, but also in relation to their fitness for forming the domains in which functions can be defined, without loss of any of the characteristic properties that serve for the classification of functions of a real variable, or of several such variables.

EXAMPLES

1. A function $f(x)$ may be defined for the interval $(0, 1)$ as follows:

$$\text{for } \frac{1}{n} \geq x > \frac{1}{n+1},\ f(x) = \frac{1}{n}x^2, \text{ and for } x = 0,\ f(0) = 1,$$

n denoting any positive integer. In this case, the norm by which the function is defined is expressible by an enumerable set of formulae which are however not independent of one another.

2. A function may be defined as follows:

$$\text{for } 1 \geqq x > \frac{1}{2},\ f(x) = \frac{x}{2};\ \text{for } \frac{1}{2} \geqq x > \frac{1}{3},\ f(x) = \frac{x}{3};\ \text{for } \frac{1}{3} \geqq x > \frac{1}{4},\ f(x) = \frac{x}{5}, \ldots,$$

and in general,

$$\text{for } \frac{1}{n} \geqq x > \frac{1}{n+1},\ f(x) = \frac{x}{P_n},$$

where P_n denotes the nth of the prime numbers 2, 3, 5, 7, ... If the function is to be defined at the point $x=0$, this may be done by assigning to $f(0)$ any arbitrarily chosen value we please. It will be observed that the values $\frac{x}{P_n}$ are in this case not representable by a single expression which involves n and x only, though they are definite single-valued functions of these.

3. Any number x, of the interval (0, 1), except 0, can be uniquely expressed in the form

$$\frac{b_1}{2} + \frac{b_2}{2^2} + \ldots + \frac{b_n}{2^n} + \ldots,$$

where b_n has, for every value of n, one of the values 0, 1, and it is stipulated that all the b_n are not to be zero from and after any fixed value of n.

A multiple-valued function* may be defined by $y = x^{\frac{1}{n}}$, where n has all positive integral values for which $b_n = 1$. This is a homonomic definition, although no analytical expression of a unitary character can be given for the representation of y.

FUNCTIONS OF A VARIABLE AGGREGATE

208. The conception of the functional relation, as it has been described above, has been restricted to that of a determinate correlation of the points of a domain, linear, plane, spatial, or p-dimensional, forming the domain of the independent variable, with a set of numbers, or points, forming the domain of the dependent variable. The modern development of the notion of the functional relation has extended it to the more general case in which a prescribed family of objects of any specified kind takes the place of the field of the independent variable. If $\{O\}$ denotes a family, or aggregate, of objects defined in accordance with some norm, and if we have also a norm by means of which a definite number is made to correspond to each member O, of the family, a variable y, identifiable with each such number, is regarded as a function of the objects O of the given family. The ordinary definition considered above of a function of a variable x, or of a number p of variables, is the particular case of the more general conception of a function, which arises when the objects O are single numbers, or points, belonging to a prescribed linear, or p-dimensional, domain. Another case of the functional relation which has recently become of considerable importance in Analysis is that in which each of the objects O consists of an enumerable set of numbers $x^{(1)}, x^{(2)}, \ldots x^{(n)}, \ldots$. Thus the functions considered may be regarded as functions of a point in space of an infinite number of dimensions; the field of the independent variable being any specified domain in such space. The particular case in which the numbers

---

* Brodén, *loc. cit.* p. 4.

$x^{(1)}, x^{(2)}, \ldots x^{(n)}, \ldots$ are so restricted that the sum of their squares forms a convergent series, in which case the domain of the point is said to be in Hilbertian space, has recently been applied in connection with the theory of integral equations. A theory of functions in Hilbertian space can be developed in this connection.

In arithmetical analysis the most general conception of a function is that in which the independent variable has as its field a family of sets of points, either linear, plane, spatial, or p-dimensional Having given, by means of an adequate definition, a family of sets of points, either discrete or continuous, a function of the sets is defined by a norm which assigns a definite number to each set of points of the given family. An important case is that in which the sets of points are all measurable; the measure of a set, as defined in accordance with the theory developed in Chapter III, is then a function of the variable set in a given family of such measurable sets.

In connection with the modern theory of integration, more general functions of measurable sets of points will emerge. The theory of functions of curves of prescribed families, developed by Volterra and others, comes under this head.

The Calculus of Variations deals with the theory of maxima and minima of functions of continuous sets of points; the functions taking the analytical form of integrals which determine a set of numbers corresponding to the continuous sets of points belonging to prescribed families.

In the present work the ordinary case of functions of one variable, or of a finite number of variables, will, for the most part, be the only one dealt with, but incidentally, as in the case of the theory of integration, functions of sets of points will be considered.

The present Chapter is concerned, in the first instance, with the properties of functions of a single variable, of which the domain is an interval, or other linear set of points. The case of functions of two or more variables, *i.e.* of points belonging to a domain in plane, or higher dimensional, space, will also be considered; especially in those respects in which the properties of such functions are not an immediate extension of the properties of functions of a single variable.

THE UPPER AND LOWER BOUNDARIES AND LIMITS OF FUNCTIONS

209. A function $y = f(x)$, being defined for the domain of x, we have seen that the values of y form a set of points, determined as usual upon a linear interval, which is called the functional image of that set of points which forms the domain of x. In case the set of points, which represent the values of y, is a bounded set, the function $f(x)$ is said to be *bounded* in the domain of x.

When the set of values of y is bounded, either boundary may be a limiting point, or only an extreme point, of the set. For convenience, and in accordance with usage, the terms "upper limit" and "lower limit" will be applied to denote the upper and lower boundaries of the derivative of this set. Thus the upper and lower limits are the extreme limiting points of y in the domain of x. Accordingly we may say that:

If the set of points y, which represents the functional image of a function $f(x)$, defined for a given domain of x, have an upper and a lower boundary, then the function $f(x)$ is said to be a bounded function, and the boundaries are said to be the upper and lower boundaries of $f(x)$ in the domain of x. The upper and lower limits of the function are the upper and lower boundaries of the limiting points of the set of points y.

The upper limit of a function $f(x)$ for its given domain may, or may not, be *attained*, *i.e.* there may, or may not, be a value of x, in the domain of x, for which the functional value is equal to the upper, or to the lower, limit. This is the case whether, or not, the upper, or the lower, limit is identical with the upper, or the lower, boundary. An upper, or a lower, boundary that is not identical with the upper, or the lower, limit, must be attained.

In case y have no upper boundary, or no lower boundary, for the domain of x, the function $f(x)$ is said to be *an unbounded function*. In this case there exist values of the function, of one sign, or of both signs, which are numerically greater than any arbitrarily assigned number A.

When y has no upper boundary in the domain of x, the function is said to have the improper boundary $+\infty$, in the domain of x. Similarly, when y has no lower boundary, it is said to have the improper boundary $-\infty$. It is frequently said, for the sake of brevity, that the upper or the lower boundary of the function is infinite.

The excess of the upper boundary of a function, in its domain, over its lower boundary, is called the fluctuation (Schwankung) *of the function in the domain.*

In case the upper or the lower boundary is infinite, the function is said to have an infinite fluctuation in its domain.

Instead of the whole of the domain of x, we may consider that part which lies in a given interval (a, b), closed or open, and the preceding definitions may be applied to this portion of the domain; thus:

The upper boundary of a function $f(x)$ in an interval (a, b), closed or open, is the upper boundary of the function when only those points of the domain of x which lie in the interval (a, b) are taken into account. A similar definition applies to the case of the upper limit, and also to the cases of the lower boundary and the lower limit.

The excess of the upper boundary of $f(x)$, in the closed, or open, interval (a, b), over its lower boundary in that interval, is called the fluctuation of $f(x)$ in the closed, or open, interval (a, b).

In case one, or both, of the boundaries is infinite, the fluctuation of the function in (a, b) is said to be infinite

If the upper boundary of $f(x)$ in (a, b) is attained, *i.e.* if there exists a value c of x such that $f(c)$ is the upper boundary, where c is a point of the domain in (a, b), then this upper boundary is said to be the *upper extreme* of the function in (a, b); and a similar definition applies to the *lower extreme.*

If the end-points a, b of the interval be left out of account, in case they belong to the domain of x, the fluctuation is called the fluctuation in the open interval (a, b). This is sometimes spoken of as the *inner fluctuation* of the function in (a, b), and is determinable as the limit of the fluctuation in the interval $(a + \epsilon, b - \epsilon)$, when ϵ is indefinitely diminished.

Precisely similar definitions are applicable to the case of a function of p variables; a cell (a, b) taking the place of an interval (a, b).

210. In accordance with the definition which has been given for a function in any domain, the value of the function at any particular point of the domain has a definite finite value. It may happen that a point P, of the domain of x, is such that, in any arbitrarily small neighbourhood of P, either the upper, or the lower, boundary of the function, or both, may not exist; so that, in any neighbourhood of P however small, there exist functional values numerically greater than any number that may be assigned. In that case, the point P is said to be *an infinity, or point of infinite discontinuity of the function*; although the function has a definite finite value at the point P itself.

Although $f(x)$ is not properly defined at a point $P(x_0)$, unless a definite numerical value be assigned to $f(x_0)$, nevertheless an improper definition of the functional value at the point P is sometimes admitted, of the form $\frac{1}{f(x_0)} = 0$; in this case the function is said to possess an *infinity* at P. This infinite discontinuity is said to be *removable*, provided that, when the functional value at P is altered to some finite value, the function have finite upper and lower boundaries in a sufficiently small neighbourhood of P.

There are other cases in which an improper definition of the functional value at a point x_0 of the domain of x is admitted. The function may be defined by means of an infinite series, of which the terms are given functions of x. This series may diverge at the particular point x_0; but it is nevertheless frequently convenient to regard the series as defining the function for all values of x in some interval which includes x_0. The functional value at x_0 is then regarded as infinite, $+\infty$, or $-\infty$.

In accordance with strict arithmetic theory, the function is regarded as undefined at points where no definite finite value of the function is specified. For the most part, in the theory which will be developed here, this restriction will be rigorously adhered to. It will be found, however, that in cases, such as in the theory of infinite series, in which it is convenient to admit improper definitions of functions at particular points, no essential change in the main results of the theory will be necessary.

In some cases it will be found convenient to remove the restriction that at each point of the domain of the independent variable the function shall be single-valued, and to define the function in such a manner that, at single points, or at each point of some set belonging to the domain of x, the function may possess finite, or infinite, multiplicity. It will be found, in the cases in which it is convenient to make this extension of the meaning of a function, that no difficulty arises as regards the use of results primarily applicable to functions which are single-valued at all points of the domain of the variable, without exception.

THE CONTINUITY OF FUNCTIONS

211. Let the domain of the independent variable x be continuous, and either bounded or unbounded; and denote the function y, at the point x, by $f(x)$. When the domain of x consists of the points of a continuous interval (a, b), that interval will be taken as closed unless the contrary is stated.

The function $f(x)$ is said to be continuous at the point a of the domain of x, if, corresponding to any arbitrarily chosen positive number ϵ whatever, a positive number δ, dependent on ϵ, exists, such that

$$|f(a+\eta)-f(a)|<\epsilon,$$

for all positive or negative values of η which are numerically less than δ, and which are such that $a+\eta$ is in the domain of x. At an end-point of a limited domain, the values of η will have one sign only.

In accordance with this definition, a neighbourhood $(a-\delta, a+\delta)$ of the point a exists, such that the function, at any point in the interior of this interval, differs numerically from its value at a, by less than ϵ. It follows that the inner fluctuation of the function in $(a-\delta, a+\delta)$ is less than 2ϵ, and it is obvious that the fluctuation in any interval interior to $(a-\delta, a+\delta)$ is less than 2ϵ. The condition of continuity of the function $f(x)$ at the point a may thus be stated to be that *a neighbourhood of the point can be determined in which the fluctuation of the function is as small as we please.*

The above definition of continuity at a point is that due to Cauchy, and is a particular case of the definition of continuity for a function of any number of variables. If we denote by $f(x^{(1)}, x^{(2)}, \dots x^{(p)})$ a function of the variables $x^{(1)}, x^{(2)}, \dots x^{(p)}$, defined for any continuous domain, the condition of continuity at the point $(\alpha^{(1)}, \alpha^{(2)}, \dots \alpha^{(p)})$ is that, corresponding to every arbitrarily chosen positive number ϵ, a number δ, dependent on ϵ, can be found, such that

$$|f(\alpha^{(1)} + h^{(1)}, \alpha^{(2)} + h^{(2)}, \dots \alpha^{(p)} + h^{(p)}) - f(\alpha^{(1)}, \alpha^{(2)}, \dots \alpha^{(p)})| < \epsilon,$$

provided $h^{(1)}, h^{(2)}, \dots h^{(p)}$ have any values which are numerically less than δ. In this case, a neighbourhood $(\alpha - \delta, \alpha + \delta)$, of a point of a linear domain, is replaced by an equilateral cell, which is a square in the case of a two-dimensional domain.

The definition of continuity has been stated by Heine* in a form which depends upon the notion of a convergent sequence of numbers or of points. Let $(P_1, P_2, \dots P_n, \dots)$ be a convergent sequence of points in the given domain, and of which P is the limiting point. The condition of continuity of the function at P is that, for every such convergent aggregate which has P as limiting point, the numbers $f(P_1), f(P_2), \dots f(P_n), \dots$ form a convergent sequence which represents the number $f(P)$. It is easily seen that a function which is continuous in accordance with Cauchy's definition is also continuous in accordance with that of Heine. For if the sequence $P_1, P_2, \dots P_n, \dots$ of points of the domain of x converges to $P(\alpha)$, we may define a sequence of numbers $\eta_1, \eta_2, \dots \eta_n, \dots$ corresponding to these points, such that $\eta_n \sim 0$, as $n \sim \infty$, and such that $PP_n < \eta_n$; it then follows by Cauchy's definition that, for a sufficiently large value of n,

$$|f(x) - f(\alpha)| < \epsilon,$$

for all the values of x which correspond to the points $P_n, P_{n+1}, \dots$. Since this holds for all values of ϵ, we see that the sequence $\{f(P_n)\}$ converges to $f(\alpha)$.

That a function $f(x)$, which is continuous at the point α, in accordance with Heine's definition, is also continuous in accordance with that of Cauchy, cannot† be proved in the general case without the employment of Zermelo's axiom (see § 197). To establish this equivalence, with the assumption of the axiom, let $H^{(\epsilon)'}(\eta)$ denote the set of points within the interval $(\alpha - \eta, \alpha + \eta)$ for which $|f(x) - f(\alpha)| \geqq \epsilon$. Assuming that there exists a set of points $H^{(\epsilon)}(\eta)$, of the domain of x, with ϵ fixed, however small η may be, (which will be the case if Cauchy's definition is not satisfied), we may consider a monotone sequence $\{\eta_n\}$ of values of η converging to zero. It is impossible that the sets $H^{(\epsilon)}(\eta_n)$ should be all identical, from and after a fixed value of n, for the intervals $(\alpha - \eta_n, \alpha + \eta_n)$

---

* *Crelle's Journal*, vol. LXXIV (1872), p. 182.

† See Sierpiński, *Comptes Rendus*, Paris, vol. CLXIII (1916), p. 688.

have no points common to an infinite number of them, except the point a which does not belong to the sets. We may consequently, without loss of generality, assume that the sets $H^{(\epsilon)}(\eta_n)$ are all different from one another. Considering the sets

$$H^{(\epsilon)}(\eta_1) - H^{(\epsilon)}(\eta_2),\ H^{(\epsilon)}(\eta_2) - H^{(\epsilon)}(\eta_3),\ \ldots\ H^{(\epsilon)}(\eta_n) - H^{(\epsilon)}(\eta_{n-1}), \ldots,$$

a set of points $p_1, p_2, \ldots p_n, \ldots$ exists in accordance with the axiom, such that p_n belongs to $H^{(\epsilon)}(\eta_n) - H^{(\epsilon)}(\eta_{n+1})$, for every value of n. We have now $|f(p_n) - f(a)| \geqq \epsilon$, for every value of n; and thus a sequence $\{p_n\}$ of points converging to a exists, such that $\{f(p_n)\}$ does not converge to $f(a)$, and hence Heine's definition of continuity at the point a is not satisfied. It follows that, if Heine's definition is satisfied, the set $H^{(\epsilon)}(\eta)$ cannot exist for all values of η; and thus that for a sufficiently small value of η, Cauchy's condition $|f(x) - f(a)| < \epsilon$ is satisfied if $|x - a| < \eta$, for all points x, of the domain of x.

A function which is not continuous at a point a may satisfy the condition that, in a neighbourhood of a, on the right the fluctuation of the function is as small as we please when the neighbourhood is small enough: the function is then said to be *continuous on the right at* a. A similar definition applies to continuity on the left.

A function is said to be *continuous in the interval* (a, b) if it satisfies the condition of continuity at every point in the interval.

For a function that is continuous in the closed interval (a, b), the upper boundary is identical with the upper limit, and the lower boundary is identical with the lower limit of the function in the closed interval. This does not necessarily hold for an open interval.

212. The domain of the independent variable has hitherto been considered to be continuous, it is however clear from a consideration of the definition of continuity, either in Cauchy's or in Heine's form, that the definition is applicable in case the domain of the independent variable is not continuous, but consists of any set of points which contains limiting points that belong to the set. It is, of course, only at such a limiting point that the question of continuity arises; for, at an isolated point of the aggregate, there are no values of the function, other than that at the point itself, in any sufficiently small neighbourhood of the point. If P be a point of the domain of x which is a limiting point of the domain, the function is continuous at P when Cauchy's definition of continuity is satisfied, those points only, in any neighbourhood of P, being taken into account, which belong to the domain for which the function is defined. If the function be continuous at every limiting point of the domain of x it is said to be continuous relatively to the given domain; and thus the notion of continuity of a function is applicable whatever be the domain of the independent variable, except when it consists of an isolated set of points.

Let $P_1, P_2, P_3, \dots$ be a convergent sequence of points of the domain of x, of which P_ω is the limiting point; and let P_ω also belong to the domain of x. Supposing the functional image, corresponding to $f(x)$, to contain the points $Q_1, Q_2, Q_3, \dots$ which correspond to $P_1, P_2, P_3, \dots$, let $Q_1, Q_2, Q_3, \dots$ form a convergent sequence of which the limiting point Q_ω corresponds to P_ω. If this condition be satisfied, however the convergent sequence be chosen in the domain of x, the aggregate (Q), of values of y, is said to be a continuous functional image of the domain (P) of x.

It is clear that the continuous functional image of a closed domain is itself closed. For there corresponds to the points of a convergent sequence $(Q_1, Q_2, Q_3, \dots)$, in (Q), an aggregate $(P_1, P_2, P_3, \dots)$, in (P), which must have at least one limiting point, and all such limiting points belong to the domain (P), and must correspond to the limiting point of $(Q_1, Q_2, Q_3, \dots)$, which therefore belongs to the aggregate (Q). Moreover if (P) be perfect, the continuous functional image (Q) is perfect also; for, corresponding to any particular point Q', of (Q), we may take a point P', of (P), for which Q' is the image. P' is the limiting point of a convergent sequence of points of (P), and to this convergent sequence there corresponds a convergent sequence in (Q), of which Q' is the limiting point. It has thus been shewn that (Q) contains no isolated points, and therefore (Q) is perfect.

If (Q) be a continuous functional image of the closed set (P), and if only one point of (P) correspond to each point of (Q), then (P) is a continuous functional image of (Q).

To the points of any convergent sequence $(Q_1, Q_2, \dots)$, in (Q), of which Q_ω is the limiting point, there corresponds a convergent sequence

$$(P_1, P_2, P_3, \dots),$$

in (P), of which P_ω is the limiting point; and P_ω is the functional image of Q_ω.

213. The theorem has been given by Weierstrass that, *if (a, b) be any interval containing points of the domain of a function, then one point at least exists in the interval, which is such that, in any arbitrarily small neighbourhood of that point, the upper boundary of the function is the same as the upper boundary of the function in the whole interval (a, b).*

This theorem holds for all functions without restriction, and it makes no difference whether the whole interval (a, b), or only a set of points in that interval, belongs to the domain of the independent variable.

Let a system $\{D_n\}$ of nets with half-closed meshes be fitted on to the interval (a, b). If M be the upper boundary of the function in that part of the domain of the independent variable contained in the closed interval

a, b), it is clear that in none of the meshes d_1, of D_1, can the upper boundary of the function be greater than M, and that in one at least of these meshes the upper boundary of the function must be M. Take the mesh d_1 of lowest rank in D_1 for which this is the case. In one at least of the meshes of D_2 that are contained in d_1, the upper boundary of the function must be M; if there are more meshes than one that satisfy this condition, take that one which is of lower rank than the others, and let this mesh be d_2. Proceeding in this manner, we obtain a sequence of meshes $d_1, d_2, \dots d_n, \dots$ each containing the next, and such that in each of them the upper boundary of the function for all the parts of the domain in the mesh is M. The single point P defined by this sequence of meshes satisfies the prescribed condition; for any neighbourhood Δ, of P, will contain all the meshes d_n, from and after some value of n; therefore in this neighbourhood the upper boundary of the function is M.

A similar result holds for the lower boundary of a function.

It is clear that this proof can be applied, in the case of functions of more variables than one, to prove the corresponding theorem that one point exists in a domain such that in its arbitrarily small neighbourhood the upper boundary of such a function is the same as in a cell.

In the case of a function which is continuous in the interval (a, b), it follows from the foregoing theorem that the upper and lower boundaries of the function in (a, b) are both finite, and thus that *a function which is continuous in an interval is bounded in that interval.*

For consider that point x_1, in (a, b), in the arbitrarily small neighbourhood of which, $(x_1 - \epsilon, x_1 + \epsilon)$, the upper boundary has the same value as for the whole interval (a, b). Since the function is continuous at x_1, corresponding to a given number δ, a number ϵ can be determined such that $|f(x) - f(x_1)| < \delta$, provided x lies in $(x_1 - \epsilon, x_1 - \epsilon)$, consequently the upper boundary of $f(x)$ in this interval must be finite, and hence $f(x)$ has a finite upper boundary in (a, b). It may be shewn in a similar manner that the function has a finite lower boundary.

A function which is continuous in the closed interval (a, b) is such that its upper limit and its lower limit are each actually attained at one point at least in the interval.

For suppose, if possible, that $f(x_1)$ has a value A, different from M: and consider an arbitrarily small interval $(x_1 - \epsilon, x_1 + \epsilon)$ for which M is the upper limit of the values of the function; then points exist in this interval for which the function differs by less than an arbitrarily small number δ, from M. These values of the function would differ from $f(x_1)$ by an amount which is not arbitrarily small, and this would be inconsistent with the condition of continuity of the function at the point x_1.

It follows that we must have $f(x_1) = M$. Similarly it may be shewn that the lower limit m is reached at least once in the interval (a, b).

A function that is continuous in an open interval (a, b) may have no upper, or no lower, boundary.

CONTINUOUS FUNCTIONS DEFINED FOR A CONTINUOUS INTERVAL

214. It will now be shewn that, *if $f(x)$ be continuous in the closed interval (a, b), and if $f(a)$, $f(b)$ have opposite signs, then there is at least one value of x in the interval, for which $f(x)$ vanishes.*

Let us suppose that $f(a) < 0$, $f(b) > 0$; and let a system of nets be fitted on to the interval (a, b). Since $f(b) > 0$, in a sufficiently small neighbourhood $(b - \epsilon, b)$ of the point b, $f(x)$ must be everywhere positive. If n be sufficiently large, one or more meshes of D_n are interior to $(b - \epsilon, b)$, and in these meshes $f(x)$ is positive. There exists consequently a mesh d_n of lowest rank in D_n, for which $f(x)$ is negative at the left hand end-point, and such that $f(x)$ is positive at the right hand end-point; unless $f(x) = 0$ at the left hand end-point, in which case a point such as the theorem requires is determined; we assume that this latter case does not arise. There is then among those meshes of D_{n+1} which are contained in d_n, one of lowest rank, such that $f(x)$ is negative, or zero, at its left hand end-point, and is positive at its right hand end-point. Proceeding in this manner, we obtain a sequence $d_n, d_{n+1}, \dots d_{n+m}, \dots$ of meshes, each of which contains the next, and such that, either, for some value of m, $f(x)$ is zero at the left hand end-point of d_{n+m}, or else, such that, for the unending sequence of values of m, $f(x)$ is negative at the left hand end-point of d_{n+m}, and is positive at the right hand end-point. In this latter case, there exists a point c, in all the meshes $\{d_{n+m}\}$. If $f(c)$ were negative, in a sufficiently small closed neighbourhood of c, $f(x)$ would be everywhere positive; and this is not possible, because d_{n+m} would be interior to that neighbourhood, provided m were sufficiently large. In a precisely similar manner, it is seen that $f(c)$ cannot be positive. Therefore $f(c) = 0$, and thus the theorem is established.

From this theorem we can deduce that, *whatever values $f(a)$, $f(b)$ may have, there must be in the interval (a, b) at least one value of x, for which $f(x)$ has any prescribed value lying between $f(a)$ and $f(b)$.*

Let this value be C, and suppose $f(a) < C < f(b)$; then the function $f(x) - C$ is continuous in the given interval, is negative when $x = a$, and positive when $x = b$; thus it vanishes at least once in the interval (a, b).

A continuous function has frequently been defined as a function such that, if $f(a)$, $f(b)$ be its values at any two points a and b, then the function passes through every value intermediate between $f(a)$ and $f(b)$, as x changes from a to b. The property contained in this definition has been shewn above to hold of every function which is continuous in accordance with Cauchy's definition; but the converse theorem does not, in general, hold. The definition just referred to is accordingly not equivalent to that of Cauchy, which is here adopted as the basis of the treatment of continuous functions. As an example of the non-equivalence of the two definitions, we may consider the function defined by $y = \sin\frac{1}{x}$, for $x \lessgtr 0$, and by $y = 0$, for $x = 0$. For this function there are values of x between a (< 0) and b (> 0) for which $f(x)$ has any assigned values c lying between $f(a)$ and $f(b)$; but the function is not continuous, in accordance with Cauchy's definition, in any interval (a, b) which contains the point 0. It is, in fact, easily seen that the point 0 is a point of discontinuity of the function; for an arbitrarily small neighbourhood of the point 0 contains points at which the function has all values in the interval $(-1, 1)$.

As another example* of a function which satisfies the condition referred to, but is discontinuous in accordance with Cauchy's definition, let the number x in the interval $(0, 1)$ be expressed as a decimal

$$\cdot a_1 a_2 a_3 \ldots a_n \ldots;$$

then consider the decimal $\cdot a_1 a_3 a_5 a_7 \ldots$. If this last decimal is not periodic, we take $f(x) = 0$; if it is periodic, and the first period commences at a_{2n-1}, we take $f(x) = \cdot a_{2n} a_{2n+2} a_{2n+4} \ldots$. The function so defined for the interval $(0, 1)$, of x, has every value between 0 and 1, in every arbitrarily small interval in the domain of x; thus the function is discontinuous at every point. A value of x for which $f(x)$ has any prescribed value $\cdot p_1 p_2 \ldots p_n \ldots$ is

$$\cdot a_1 A a_3 B \ldots K a_{2n-1} p_1 a_{2n} p_2 a_{2n+2} \ldots,$$

where $\cdot a_1 a_3 \ldots$ is any periodic decimal, the first period of which begins at a_{2n-1}, and $A, B, \ldots K$ are arbitrarily chosen digits. Nevertheless there are values of x, between α and β, at which the function takes any assigned value intermediate between $f(\alpha)$ and $f(\beta)$.

CONTINUOUS FUNCTIONS DEFINED AT POINTS OF A SET

215. It will now be shewn that, *if a function $f(x)$, having prescribed values at each point of an infinite set of points in the interval (a, b), be continuous in that interval, then the values of the function are determinate at each point of the derivative of the set.*

Suppose $\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n, \ldots$ to be a convergent sequence of points, for which x_1 is the limiting point, and suppose $f(\alpha_1), f(\alpha_2), \ldots f(\alpha_n), \ldots$ to be known; it will be shewn that these functional values form a convergent sequence whose limit is $f(x_1)$. An interval $(x_1 - \delta, x_1 + \delta)$ can always be found, corresponding to any fixed number ϵ, such that the function at any point of this interval differs from $f(x_1)$ by less than the arbitrarily small number ϵ; this follows from the continuity of the function. A number n can be found such that all the points $\alpha_n, \alpha_{n+1}, \alpha_{n+2}, \ldots$ lie within the interval $(x_1 - \delta, x_1 + \delta)$. It follows that

---

* See Lebesgue, *Leçons sur l'intégration*, p. 90.

$$|f(x_1) - f(\alpha_n)| \text{ and } |f(x_1) - f(\alpha_{n+1})|,$$

&c. are all less than ϵ, which is arbitrarily small; hence $f(x_1)$ is the limit of the sequence $f(\alpha_1), f(\alpha_2), \ldots$, and thus $f(x_1)$ is determinate. From this special case it follows that, for all the limiting points of a given set of points in (a, b), the values of the continuous function are determinate. It thus appears that the function is determinate for all points which belong to the derivative of the given set, for the points of which the values of the function are known.

In particular, *if a continuous function have prescribed values for points of a set which is everywhere dense throughout the interval* (a, b), *then its values are determinate for all points of the interval.*

A special case of such a set would be that of the rational points within the interval. It follows that a continuous function whose values are known for all the rational points in an interval is determinate for all the irrational points. A continuous function which is known to be constant for all the rational points has the same constant value for all the irrational points in the interval.

It is clear that, if cells be employed, instead of intervals, the method in the proof is applicable to shew that a continuous function of any number of variables is determinate at each point of the derivative of a set of points at which the values of the continuous function are assigned.

A generalization of the above theorem is that, if a function is continuous with reference to a domain which consists of a set (P), and is known for all points of a sub-set which is everywhere dense in (P), then the function is determinate for every point of (P). This may be seen by remembering that every point of (P) is a limiting point of the sub-set, and applying the same reasoning as before.

216. From the theorem established above, that a continuous function is determinate when its values at an everywhere dense enumerable set of points are prescribed, we may deduce that *the cardinal number of the aggregate of all continuous functions of a real variable is the cardinal number* c, *of the continuum.*

We may suppose the values of a function to be prescribed at the rational points. The cardinal number of the aggregate of all functions defined for the rational points only is the cardinal number of the ways of distributing on the aggregate of rational numbers the aggregate of numbers of the continuum. This number is c^a, which has been shewn in § 183 to be equal to c. Only some of the distributions of this kind are such as will give rise to continuous functions; hence the aggregate of all continuous functions is a part of the aggregate of all possible distributions on the set of rational numbers of the numbers of the continuum. It follows that the cardinal number of the aggregate of all continuous functions is $\leq c$. Again, this cardinal number is $\geq c$; for among the continuous functions are those which are constant, and everywhere equal to any assigned number of the continuum; and thus the aggregate of all continuous functions contains a part which has the cardinal number c. Since the cardinal number is $\leq c$, and also $\geq c$, it is equal to c.

This theorem was established by Borel*, who also shewed that the aggregate of all analytical functions of two or more variables has the cardinal number c.

The cardinal number of the aggregate of all functions of a real variable is that of all distributions of the continuum upon itself; this is, in accordance with the definition in § 150, denoted by c^c, for which we may write f

Each particular distribution of the numbers of the continuum on themselves is definable by a definite norm, and corresponding to each such distribution there is a definite function for a continuous domain. Let the aggregate of all such functions be denoted by F: it will then be proved that the cardinal number f, of F, is $> c$.

First, F has a part which is equivalent to the continuum. This is at once seen, since the functions $f(x) = c$, where c is any number of the continuum, constitute such a part. It follows that $f \geq c$.

Next, let it be assumed, if possible, that F is equivalent to a part of the continuum. As has just been proved, such a part cannot have a cardinal number $< c$; we therefore assume that F is equivalent to the set of numbers of the continuum. This amounts to the assumption that F can be ordered in the same type as the continuum, so that, to any assigned number ξ of the continuum, there corresponds a definite set of rules R_ξ, which defines a function $f_\xi(x)$. The correspondence between ξ and R_ξ must itself be defined by a set of rules, so that when ξ is assigned, R_ξ, and therefore the function $f_\xi(x)$, is defined. The aggregate $\{f_\xi(x)\}$ must contain every definable function of a real variable. The number ξ being assigned, $f_\xi(x)$ is producible; and its existence implies that, at any assigned point ξ', the functional value $f_\xi(\xi')$ can be determined arithmetically. We may take, for example, $\xi' = \xi$; and thus, if ξ is assigned, $f_\xi(\xi)$ is known. We may regard $f_\xi(\xi)$ as a function of ξ; for its value at any point ξ can be arithmetically determined, and it is therefore an element of the aggregate F of all functions. With this understanding as to $f_\xi(\xi)$, choose a fixed number, say unity; then the function $\phi(\xi) \equiv f_\xi(\xi) + 1$ has a definite norm; for we have only to add to the rules by which $f_\xi(\xi)$ is defined, the further rule, that, at each point ξ, unity is to be added to the value of $f_\xi(\xi)$. We have now a new definable function $\phi(x)$; but this cannot possibly belong to the aggregate F, for if it do so belong, there must be some one point ξ_1 of the continuum, with which it corresponds: but $\phi(x)$ cannot be identical with $f_{\xi_1}(x)$, for $\phi(\xi_1)$ and $f_{\xi_1}(\xi_1)$ differ by unity. Since $\phi(\xi)$ is not contained in F, contrary to the hypothesis, it follows that F cannot be equivalent to the continuum, and thus the theorem, $f > c$, is established. It has therefore been shewn that $f > c$, and consequently that—*the aggregate of all functions of a real variable has a cardinal number f, greater than c.*

---

* See *Leçons sur la théorie des fonctions*, p. 127.

UNIFORM CONTINUITY

217. It will now be shewn that, if the domain of x be a continuous closed interval, then a continuous function is *uniformly continuous* through the domain of x. It will be proved that a number $\bar{\delta}$ can be found, corresponding to any given ϵ, such that, for all values of x, the fluctuation of $f(x)$ within the neighbourhood $(x - \bar{\delta}, x + \bar{\delta})$, or for all of this neighbourhood which lies within the domain, is less than the number ϵ. The essential point is that $\bar{\delta}$ is independent of x.

Consider a symmetrical system $\{D_n\}$, of nets with closed meshes, fitted on to the interval (a, b).

If ϵ be a prescribed positive number, there must be some smallest integer n_1, such that, in each of the meshes of the net D_{n_1}, the fluctuation of $f(x)$ is $< \frac{1}{2}\epsilon$. For suppose that no such integer n_1 exists: then, however large n may be taken, there is one mesh at least of D_n in which the fluctuation of $f(x)$ is $\geq \frac{1}{2}\epsilon$. Let D_n' be the set of all the meshes of D_n for which this is the case; the sequence of sets of meshes $D_1', D_2', \dots D_n', \dots$, each of which contains the next, defines a closed set of points in all of them. Let P be such a point, and let a neighbourhood Δ', of P, be constructed, within which the fluctuation of the function is $< \frac{1}{2}\epsilon$. This neighbourhood Δ' will contain, in its interior, a mesh of D_n', from and after some fixed value of n, and in each of these the fluctuation of $f(x)$ is $\geq \frac{1}{2}\epsilon$, and thus there is a contradiction in the assumption that such a point as P exists. Therefore D_n' cannot exist for all values of n, and thus, for some integer n_1, the fluctuation in every net of D_{n_1} is $< \frac{1}{2}\epsilon$.

If x be any point of the closed interval (a, b), and the meshes of D_{n_1} be all of breadth $2\bar{\delta}$, the interval $(x - \bar{\delta}, x + \bar{\delta})$, or the part of it that is in (a, b), is in the interval formed by at most two consecutive meshes of D_{n_1}, and the inner fluctuation of $f(x)$ in $(x - \bar{\delta}, x + \bar{\delta})$ is consequently $< \epsilon$.

It has thus been shewn* that it is unnecessary, for a function of a real variable, to draw a distinction, as has sometimes been done, between functions which are uniformly, and those which are non-uniformly, continuous in the continuous domain of x; for all continuous functions are uniformly continuous.

---

* This theorem was first stated and proved by Heine; see *Crelle's Journal*, vol. LXXI (1870), p. 361, and vol. LXXIV (1872), p. 188.

The theorem may also be stated in the following form:

If $f(x)$ be continuous in the closed interval (a, b), then, corresponding to any arbitrarily chosen positive number ϵ, a number η can be determined, such that the condition $|f(x_1) - f(x_2)| < \epsilon$, is satisfied, for any two points x_1, x_2, in (a, b), such that $|x_1 - x_2| < \eta$.

The following theorem can be immediately deduced:

If a function be continuous in a finite closed interval, then the interval can be divided into a finite number of sub-intervals, in every one of which the fluctuation of the function is less than a prescribed positive number.

It is in fact clear that, if ϵ be the prescribed number, the condition is satisfied when the interval is subdivided in any manner such that the length of the greatest of the sub-intervals is $< \eta$.

Another proof of the above theorem, in an extended form, will be given in § 234, by employing the Heine-Borel theorem.

It is clear that the above proof applies also to the case in which the domain of x is not a continuum, but is any closed set: because the essential point of the proof depends upon the limiting points all belonging to the domain. Those meshes of the nets which contain no points of the domain of x are disregarded. For domains which are not closed the proof does not apply; thus a function which is continuous, relatively to an aggregate which is not closed, is not necessarily uniformly continuous.

In the case of a function of p variables, when the domain of the variables is a closed set, the neighbourhood

$$(x_1 - \bar{\delta}, x_2 - \bar{\delta}, \dots x_p - \bar{\delta}; x_1 + \bar{\delta}, x_2 + \bar{\delta}, \dots x_p + \bar{\delta}),$$

of the point $(x_1, x_2, \dots, x_p)$, may be (§ 49) denoted by $(x - \bar{\delta}, x - \bar{\delta})$. The condition of uniform continuity takes then the same form as in the case $p = 1$. If we take the cell in which the domain is contained to have its edges all equal, and employ a system of nets in which the edges of each mesh are all equal, the above proof can be applied to shew that a function that is continuous in a domain that consists of a closed set of points is uniformly continuous.

ABSOLUTE CONTINUITY

218. *A function $f(x)$, defined in the closed interval (a, b), is said to be absolutely continuous in (a, b) if, corresponding to an arbitrarily chosen positive number ϵ, another positive number η can be so determined that, in every enumerable, or finite, set of non-overlapping intervals*

$$(x_1, x_1'), (x_2, x_2'), \dots (x_n, x_n'), \dots,$$

in the interval (a, b), *and such that the total measure of the intervals is* $< \eta$, *the sum, or limiting sum,* $\sum\limits_{r=1} | f(x_r) - f(x_r') |$ *is* $< \epsilon$.

It is clear that a function which is absolutely continuous in (a, b) is also continuous, in accordance with the definition of § 211; but the converse does not hold. We may take the set of intervals to consist of a single interval of length $< \eta$; thus the condition of uniform continuity is satisfied.

It is easily seen that an equivalent form of the definition is that, in every non-overlapping set of intervals of total measure $< \eta$, the sum, or limiting sum, $\sum\limits_{r=1} \{f(x_r) - f(x_r')\}$ is in absolute value $< \epsilon$. For we can divide the set of intervals into two parts, those for which $f(x_r) - f(x_r')$ is positive and those for which it is negative.

The sum and the product of two functions which are absolutely continuous in (a, b) *are absolutely continuous in the interval.*

Let $f_1(x), f_2(x)$ be absolutely continuous in (a, b); and let their sum be denoted by $f(x)$. We have then

$$\sum_{r=1} | f(x_r) - f(x_r') | \leq \sum_{r=1} | f_1(x_r) - f_1(x_r') | + \sum_{r=1} | f_2(x_r) - f_2(x_r') |.$$

If ϵ be arbitrarily chosen, and if the measure of the set of intervals is less than some number η_1, the first sum on the right hand side is $< \frac{1}{2}\epsilon$; also there exists a number η_2 such that the second sum on the right hand side is less than $\frac{1}{2}\epsilon$, provided the measure of the set of intervals is less than η_2. If η be the smaller of the numbers η_1, η_2, we have

$$\sum_{r=1} | f(x_r) - f(x_r') | < \epsilon,$$

provided the measure of the set of intervals is $< \eta$; therefore $f(x)$ satisfies the condition of absolute continuity.

Next, let $f(x) = f_1(x) f_2(x)$; then since

$$f(x_r) - f(x_r') = f_1(x_r) \{f_2(x_r) - f_2(x_r')\} + f_2(x_r') \{f_1(x_r) - f_1(x_r')\},$$

we have

$$\sum_{r=1} | f(x_r) - f(x_r') | \leq M_1 \sum_{r=1} | f_2(x_r) - f_2(x_r') | + M_2 \sum_{r=1} | f_1(x_r) - f_1(x_r') |;$$

where M_1, M_2 are the upper boundaries of $| f_1(x) |$, $| f_2(x) |$ in (a, b). A number η can be so chosen that

$$\sum_{r=1} | f_1(x_r) - f_1(x_r') | < \frac{\epsilon}{2M_2}, \quad \sum_{r=1} | f_2(x_r) - f_2(x_r') | < \frac{\epsilon}{2M_1},$$

provided the measure of the set of intervals is $< \eta$. When this last condition is satisfied, we have

$$\sum_{r=1} | f(x_r) - f(x_r') | < \epsilon;$$

and therefore the product $f_1(x) f_2(x)$ is absolutely continuous in (a, b).

If $f(x)$ be absolutely continuous in the interval (a, b), of x, and $\phi(t)$ be absolutely continuous in the interval (α, β), of t, then, if x and t be related by the condition $x = \phi(t)$, the function $f\{\phi(t)\}$ can easily be shewn to be absolutely continuous, in case the function $\phi(t)$ is also monotone (see § 239). It was however pointed out by Dunham Jackson* that this is not necessarily the case when the absolutely continuous function $\phi(t)$ is not monotone. De la Vallée Poussin has given a sufficient condition† to be satisfied by $\phi(t)$, in order that $f\{\phi(t)\}$ may be absolutely continuous for all absolutely continuous functions $f(x)$; and the matter has been further investigated by Fichtenholz‡. The latter has shewn that the necessary and sufficient condition that $f\{\phi(t)\}$ should be absolutely continuous, where $f(x)$ and $\phi(t)$ are absolutely continuous, is that $f\{\phi(t)\}$ should be of bounded variation (see § 243).

THE CONTINUITY OF UNBOUNDED FUNCTIONS

219. Let it be assumed that the continuous function $\phi(x)$, defined in the interval (a, b), of x, has 1 and -1 for its upper and lower boundaries, these being, in virtue of the continuity of the function, also its upper and lower limits.

Consider the function $f(x) = \dfrac{\phi(x)}{1 - |\phi(x)|}$, from which it follows that $\phi(x) = \dfrac{f(x)}{1 + |f(x)|}$; the functions $\phi(x)$, $f(x)$ have the same sign at each point x. When $\phi(x) = 1$, we may suppose the improper value ∞ to be assigned to $f(x)$; and when $\phi(x) = -1$, we may suppose the improper value $-\infty$ to be assigned to $f(x)$. In this manner, by adjoining to the real numbers the two elements ∞, $-\infty$, we ensure that a (1, 1) correspondence exists, without exception, between the values of $\phi(x)$, $f(x)$. Thus, to the closed interval $(-1, 1)$, there now corresponds the *closed* interval $(-\infty, \infty)$, and, in the correspondence so defined, the relative order of two elements is preserved without change; thus, as $\phi(x)$ increases steadily from -1 to 1, $f(x)$ increases steadily from $-\infty$ to ∞.

The function $f(x)$ is continuous for any finite value a, of x, which is not such that $\phi(a)$ has one of the values $1, -1, 0$.

For
$$f(x) - f(a) = \frac{\phi(x) - \phi(a)}{\{1 - |\phi(x)|\}\{1 - |\phi(a)|\}},$$
provided $f(x)$ and $f(a)$ have the same sign. A neighbourhood of a can be so determined that, at every point of that neighbourhood,
$$1 - |\phi(x)| > h,$$

---

* See de la Vallée Poussin, *Trans. Amer. Math. Soc.* vol. XVI (1915), p. 462, footnote.

† *Loc. cit.* p. 462.

‡ *Bulletin de l'acad. roy. de Belgique* (1922), Nos. 6, 7, p. 430.

where h is some positive number so chosen that $1 - |\phi(a)| > h$, and also such that $|\phi(x) - \phi(a)| < h^2\epsilon$, where ϵ is an arbitrarily chosen positive number. In this neighbourhood we have then $|f(x) - f(a)| < \epsilon$, and thus $f(x)$ is continuous at a.

If $\phi(a) = 0$, we have $|\phi(x)| < \epsilon$, in a sufficiently small neighbourhood of a, and therefore $|f(a)| < \frac{\epsilon}{1-\epsilon} < 2\epsilon$, if $\epsilon < \frac{1}{2}$. Hence $f(x)$ is continuous at a.

At a point a at which $\phi(x) = 1$, the function $f(x)$ has the value ∞. If N be an arbitrarily large number, a neighbourhood of a exists in which $\phi(x) > \frac{N}{N+1}$, in virtue of the continuity of $\phi(x)$, at a; so that $f(a) > N$ in that neighbourhood. In a similar manner, it may be shewn that, at a point a, at which $f(a) = -1$, a neighbourhood of a exists such that $f(a) < -N$.

It is convenient to extend the definition of continuity of a function at a point in such wise that the points at which $f(x)$ has the values ∞, $-\infty$ are points of continuity of the function, corresponding to the continuity of the function $\phi(x)$ at such points.

We are thus led to the following definition of the continuity of an unbounded function at a point in the neighbourhood of which the function has indefinitely great values:

A function $f(x)$ is said to be continuous, in the extended sense, at a point a, when, corresponding to any arbitrarily chosen positive number N, a neighbourhood of a exists, such that, at every point of that neighbourhood, $f(x) > N$. It is also said to be continuous at a, if the above conditions be replaced by $f(x) < -N$. The value of the function at a may be regarded as ∞, in the first case, and $-\infty$ in the second case.

An unbounded function $f(x)$ is regarded as continuous in the interval (a, b), for which it is defined, if it is continuous at every point $f(x)$ in the neighbourhood of which the function is bounded, and is continuous, in the extended sense, at every other point.

Thus, in the above case, the unbounded function $f(x)$, continuous in this sense, corresponds to $\phi(x)$ which is continuous in the more restricted sense of the term. It can easily be seen that the continuity of $\phi(x)$ is a consequence of that of $f(x)$.

In the above extension of the conception of a continuous function to the case of unbounded functions, the points ∞ and $-\infty$ have been regarded as distinct.

It is however possible to regard these points as not distinct from one another, so that when we consider the functions $f(x)$, $\frac{1}{f(x)}$, a value at which $f(x) = 0$ corresponds to a single value $\pm\infty$, of $\frac{1}{f(x)}$.

In this case *a function $f(x)$ is said to be continuous at a point a for which, corresponding to each arbitrarily chosen positive number N, a neighbourhood of a exists, in the whole of which $|f(x)| > N$.*

It is clear that this condition may be satisfied in cases in which the condition of continuity is not satisfied when ∞ and $-\infty$ are regarded as distinct values of a function.

THE LIMITS OF A FUNCTION AT A POINT

220. Let a be a limiting point of the set of points which forms the domain of the independent variable x; the point a may, or may not, itself belong to the domain of x. Let $(a, a + h)$ be a neighbourhood of a on the right, and let $U(h)$, $L(h)$ denote the upper and lower limits of a given function $f(x)$ for all the points of the domain of x which are interior to the interval $(a, a + h)$. It will be observed that $f(a)$, if it exists, is not reckoned amongst the functional values of which $U(h)$, $L(h)$ are the upper and lower limits.

Let a sequence of diminishing values be assigned to h, which converges to zero; denoting this sequence by $h_1, h_2, h_3, \ldots$: the corresponding numbers $U(h_1), U(h_2), \ldots U(h_n), \ldots$ form a sequence of which the members do not increase, and therefore they have in general a definite lower limit, which is called the *upper limit of $f(x)$ at a on the right*. It is easily seen that this limit is independent of the particular sequence $\{h_n\}$, employed in defining it. It may happen that all the upper limits $U(h)$ are infinite, in which case we say that the upper limit of $f(x)$ at a on the right is $-\infty$: or it may happen that the sequence $U(h_1), U(h_2), \ldots U(h_n), \ldots$ has no lower limit, in which case we say that the upper limit of $f(x)$ at a on the right is $-\infty$. In any case, the finite, or infinite, upper limit of $f(x)$ at a, on the right, is denoted by $\overline{f(a + 0)}$.

The numbers $L(h_1), L(h_2), \ldots L(h_n), \ldots$ form a sequence of which the elements do not diminish, and they have an upper limit, which is called the *lower limit of $f(x)$, at a on the right*, and may, as in the former case, have infinite values ∞, or $-\infty$. This limit is denoted by $\underline{f(a - 0)}$.

Corresponding definitions apply to the left of the point a; and the limits of $f(x)$ at a on the left are denoted by $\overline{f(a - 0)}$, $\underline{f(a - 0)}$ respectively. In case the point a is a limiting point of the domain of x, on one side only, the two limits of the function at a on the other side are non-existent.

In defining $\overline{f(\alpha + 0)}$ it is immaterial whether we employ the upper limit, or the upper boundary, of the function in the open interval $(\alpha, \alpha + h)$. If $U(h)$ be the upper limit, and $\overline{U}(h)$ the upper boundary, all except a finite number of values of $f(x)$ in the open interval $(\alpha, \alpha + h)$ are such that $f(x) < U(h) + \epsilon_h$; where ϵ_h is an arbitrarily chosen positive number. A number $h' < h$ can be so chosen that $U(h') \leqq \overline{U}(h') < U(h) + \epsilon_h$. If this be done in such a manner that $\lim\limits_{h \sim 0} \epsilon_h = 0$, it follows that $U(h)$ and $\overline{U}(h)$ have the same lower limit, as $h \sim 0$.

A similar remark applies to $\underline{f(\alpha + 0)}$, $\overline{f(\alpha - 0)}$, $\underline{f(\alpha - 0)}$.

The definitions of the upper and lower limits at a point α may be stated shortly as follows:

The upper limit $\overline{f(\alpha + 0)}$ *of a function at* α, *on the right, is the limit of the upper limit of* $f(x)$ *in the open interval* $(\alpha, \alpha + h)$, *when* h *is indefinitely diminished.*

The lower limit $\underline{f(\alpha + 0)}$ *of a function at* α, *on the right, is the limit of the lower limit of* $f(x)$ *in the open interval* $(\alpha, \alpha + h)$, *when* h *is indefinitely diminished.*

The definitions for the left of α may be stated in a precisely similar manner.

It is to be observed that the four functional limits $\overline{f(\alpha + 0)}$, $\underline{f(\alpha + 0)}$, $\overline{f(\alpha - 0)}$, $\underline{f(\alpha - 0)}$ are entirely independent of $f(\alpha)$, in case α belongs to the domain for which $f(x)$ is defined. Any arbitrary alteration in the value of $f(\alpha)$ will not affect these four limits of $f(x)$, at α.

The conditions that the point α may be a point of continuity of the function $f(x)$ are that $\overline{f(\alpha + 0)}$, $\underline{f(\alpha + 0)}$, $\overline{f(\alpha - 0)}$, $\underline{f(\alpha - 0)}$, $f(\alpha)$ should all have the same finite value.

It may happen that $f(\alpha)$, $\overline{f(\alpha + 0)}$, $\underline{f(\alpha + 0)}$ have one and the same finite value, but that either, or both, of $\overline{f(\alpha - 0)}$, $\underline{f(\alpha - 0)}$ have not this value; in this case $f(x)$ is said to be continuous at α, on the right. Continuity at α, on the left, is defined in a similar manner.

If the four functional limits at α be all finite and equal, but $f(\alpha)$ have a different value, then the function is said to have a *removable discontinuity* at the point α. In this case the function would be made continuous at α merely by properly altering the value of $f(\alpha)$.

The four functional limits at the point $x = 0$ are usually denoted by $\overline{f(+0)}$, $\underline{f(+0)}$, $\overline{f(-0)}$, $\underline{f(-0)}$ respectively.

If, at a point α,

$$f(\alpha) = \overline{f(\alpha + 0)} = \underline{f(\alpha + 0)} = \overline{f(\alpha - 0)} = \underline{f(\alpha - 0)} = +\infty,$$

or if $f(a)$, and all the four limits, are $-\infty$, the function is said to be continuous at a, in the extended sense of the term; the distinction between $+\infty$ and $-\infty$ being preserved.

In case $f(a) = \overline{f(a+0)} = \underline{f(a+0)} = +\infty$, or $-\infty$, the function is said to be continuous at a, on the right, in the extended sense of the term.

The four numbers $\overline{f(x+0)}$, $\underline{f(x+0)}$, $\overline{f(x-0)}$, $\underline{f(x-0)}$, which are dependent on x, may be regarded as defining functions of x called the upper, and the lower, right hand, or left hand, limiting functions associated with $f(x)$.

The *upper associated function* $A(x)$ may be defined to be a function whose value at x is the greater of the two numbers $\overline{f(x-0)}$, $\overline{f(x-0)}$: and the *lower associated function* $a(x)$ may be defined to be a function whose value at x is the lesser of the two numbers $\underline{f(x+0)}$, $\underline{f(x-0)}$.

The greatest of the three numbers $f(x)$, $\overline{f(x+0)}$, $\overline{f(x-0)}$ may be regarded as the value, at x, of a function $M(x)$, called *the maximal function associated with* $f(x)$.

The least of the three numbers $f(x)$, $\underline{f(x+0)}$, $\underline{f(x-0)}$ may be regarded as the value, at x, of a function $m(x)$, called *the minimal function associated with* $f(x)$.

221. If the upper and lower limits of $f(x)$ at a, on the right, have the same value, this common value is called *the limit of* $f(x)$ *at* a, *on the right*, and is denoted[*] by $f(a+0)$. If the upper and lower limits of $f(x)$ at a, on the left, have the same value, this is called the limit of $f(x)$ at a, on the left, and is denoted by $f(a-0)$. Either of the limits, on the right, or left, at a point, when such limit exists, may be either finite or infinite

The limit at $x = 0$, on the right, is denoted by $f(+0)$: and the corresponding limit on the left is denoted by $f(-0)$.

The limit at a point P, on one side, may be also defined as follows:—Let $(P_1, P_2, P_3, \ldots)$ be any convergent sequence of points belonging to the domain of x, which is such that a, or P, is its limiting point, and such that all the points of the sequence are on the one side of P. The values of $f(x)$ at $P_1, P_2, P_3, \ldots$ form an aggregate which may be a convergent sequence; let us suppose it to be so, and also that its limit has a value which is independent of the particular sequence, which is however subject to the conditions above stated. In that case, this limit is denoted by $f(a+0)$, or by $f(a-0)$, as the case may be, and is called the limit of $f(x)$ at a, on the right or left.

It may be observed that the necessary and sufficient condition for the existence of a definite finite limit, on the right, at a, is that, corresponding to every arbitrarily small number ϵ, a neighbourhood $(a, a + \delta)$ can be found, such that the difference of the values of the function at every pair of points of the domain of x, which are in the interior of this interval, is numerically less than ϵ.

---

[*] This notation was introduced by Dirichlet; see *Werke*, vol. I, p. 156.

The necessary and sufficient condition that $f(a + 0)$ should exist, and be equal to $+\infty$, is that, if A be an arbitrarily chosen positive number, δ can be so determined that at every point interior to $(a, a + \delta)$ the condition $f(x) > A$ is satisfied. In order that $f(a + 0)$ may exist, and be equal to $-\infty$, the corresponding condition is that $f(x) < -A$.

It is possible that one of the limits $f(a + 0)$, $f(a - 0)$ may exist and not the other. If the domain of x be either a continuous interval, or a perfect set of points, a may be taken to be at any point of the domain.

When the condition for the existence of $f(a + 0)$, or of $f(a - 0)$, at a point a, is not satisfied, the convergent sequence $(P_1, P_2, \ldots P_n, \ldots)$, of which $P(a)$ is the limiting point, may be such that

$$f(P_1), f(P_2), \ldots f(P_n), \ldots$$

is either not a convergent sequence, or else that its limit depends upon the particular choice of the points $P_1, P_2, \ldots P_n, \ldots$. In this case the fluctuation of $f(x)$ within an arbitrarily small neighbourhood $(a, a + \delta)$, on the one side of a, is either a finite number which has not zero for its limit, when δ is indefinitely diminished, or else it is indefinitely great, however small δ may be.

222. If $x_1, x_2, x_3, \ldots x_n, \ldots$ be a convergent sequence of points belonging to the domain of x, with a for its limiting point, then the sequence $f(x_1), f(x_2), \ldots f(x_n), \ldots$ may not be convergent; but, if it be convergent, its limit may have either (1), a single value independent of the mode in which the convergent sequence is chosen, in which case a is either a point of continuity of $f(x)$, or a point of removable discontinuity of the function; or (2), one of two values, in which case both the limits $f(a + 0)$, $f(a - 0)$ exist; or (3), one of a finite, or an indefinitely great, number of values, which all lie between the greatest and least of the four functional limits at a.

If the sequence $f(x_1), f(x_2), \ldots f(x_n), \ldots$ be not convergent, by omitting a certain set of the numbers of the sequence, we shall obtain a convergent sequence, or else one which diverges to $+\infty$, or to $-\infty$. We can therefore without loss of generality suppose that $x_1, x_2, \ldots x_n, \ldots$ are so chosen that this is the case.

The aggregate of all possible values of the limits of such a sequence $f(x_1), f(x_2), \ldots f(x_n), \ldots$, for all sequences $\{x_n\}$ which converge to a, each of which limits has a finite value, or is ∞, or $-\infty$, is called the ***aggregate of functional limits*** at the point a. Of the numbers $x_1, x_2, \ldots x_n, \ldots$ an infinite number are on one side of a; we may therefore without loss of generality assume that $x_1, x_2, \dots x_n, \dots$ are all on the same side of a.

It will be shewn that *the aggregate of functional limits at a is a closed set* (which may be finite), provided, in case any functional limit is infinite, we regard the point ∞, or $-\infty$, as belonging to the set.

Suppose U is the limit to which such a sequence converges. If the aggregate of all values of U is not finite, let $U_1, U_2, \dots U_n, \dots$ be a convergent sequence of values of U, of which U_ω is the limit. It will be shewn that U_ω is itself a value of U. First suppose that U_ω is finite. Let U_r be the limit of a sequence $\{f(x_n^{(r)})\}$, for $n = 1, 2, 3, \dots$; we may choose r so large that

$$| U_\omega - U_r | < \tfrac{1}{2}\epsilon,$$

for this value, and all greater values, of r. We can then choose n so that

$$| U_r - f(x_n^{(r)}) | < \tfrac{1}{2}\epsilon,$$

for that value, and all greater values, of n. It follows that for all sufficiently large values of r and n, $| U_\omega - f(x_n^{(r)}) | < \epsilon$.

As ϵ is arbitrarily small, we can obtain a sequence of numbers x such as $x_n^{(r)}$, for which $f(x_n^{(r)})$ converges to U_ω, whilst the numbers $x_n^{(r)}$ converge to a. Hence U_ω belongs itself to the aggregate of functional limits.

In case U_ω is infinite, we can choose r so that $| U_r | > A$, an arbitrarily chosen number; then, as before, n can be so chosen that

$$| U_r - f(x_n^{(r)}) | < \tfrac{1}{2}\epsilon,$$

and thus $| f(x_n^{(r)}) | > A - \tfrac{1}{2}\epsilon$. Taking an increasing sequence of values of A, and a diminishing sequence of values of ϵ, we obtain a sequence of points $x_n^{(r)}$, for which $| f(x_n^{(r)}) |$ is divergent.

It thus appears that the aggregate of functional limits at a point a is closed, in the ordinary sense, when the upper and lower functional limits, at a, are all finite; and that, when this is not the case, the set will be closed if we regard one, or both, of the points ∞, $-\infty$ as belonging to it.

The aggregate of functional limits may be finite, or may consist of a closed set, of any type as regards density.

223. If the domain of x be unbounded, in one, or in both, directions, it may happen that a point x_1 (> 0) of the domain can be found, corresponding to every arbitrarily chosen positive number ϵ, such that the difference between the values of $f(x)$, for any two values of x which are both greater than x_1, is numerically less than ϵ. In this case the function has a definite limit, as x is increased indefinitely in its domain; and this is called the limit of $f(x)$ for $x \sim \infty$. Under a corresponding condition $f(x)$ may have a definite limit for $x \sim -\infty$.

If, as x increases, a point x_1 of the domain of x, corresponding to each assigned positive number A, chosen as great as we please, can be found, such that $f(x) > A$, for all values of x which belong to the domain, and are $> x_1$, then the limit of $f(x)$ is said to be ∞, as x is increased indefinitely. If $f(x) < -A$, for all such values of x_1, then the limit of $f(x)$ is said to be $-\infty$. Similar definitions apply to the case in which x has indefinitely great values in the negative direction.

In case the limit $f(\alpha+0)$, at a point α on the right, do not exist as a definite number, and be not infinite with a fixed sign, it is frequently convenient to regard $f(\alpha+0)$ as still existent, but indeterminate, and capable of all values belonging to some closed set of which $\overline{f(\alpha+0)}$, $\underline{f(\alpha+0)}$ are the extreme values. It is then said that $f(\alpha+0)$ is indefinite in value, and that $\overline{f(\alpha+0)}$, $\underline{f(\alpha+0)}$ are its *limits of indeterminancy*. A similar remark applies to $f(\alpha-0)$, which may also be either definite, or indefinite, with $\overline{f(\alpha-0)}, \underline{f(\alpha-0)}$ as its limits of indeterminancy. One, or both, of the limits of indeterminancy, in either case, may be infinite.

THE DISCONTINUITIES OF FUNCTIONS

224. Let us suppose the domain of x to include all points in a sufficiently small neighbourhood of a point α; or, in any case, let α be a limiting point of the domain of x.

The fluctuation of the function $f(x)$ in the closed, or open, neighbourhood $(\alpha-\delta, \alpha+\delta)$ of the point α, depends in general upon δ, but cannot increase as δ is diminished. It therefore has a lower limit, for values of δ which converge to zero. This limit, which may be zero, finite, or indefinitely great, is called the saltus (Sprung), or measure of discontinuity, of the function $f(x)$, at α; thus:

The saltus, or measure of discontinuity, of a function $f(x)$, *at a point* α, *is the limit of the fluctuation of the function in a neighbourhood* $(\alpha-\delta, \alpha+\delta)$, *as* δ *converges to zero.*

The upper boundary of the function $f(x)$ *in the interval* $(\alpha-\delta, \alpha+\delta)$ *has a lower limit, as* δ *is indefinitely diminished, which is called the maximum* $M(\alpha)$, *of the function* $f(x)$, *at* α (see § 220).

The lower boundary of the function, in the same interval, has an upper limit, as δ *is indefinitely diminished, which is called the minimum* $m(\alpha)$, *of* $f(x)$, *at* α. Either the maximum or the minimum at a point may be indefinitely great.

The saltus of $f(x)$ at α is easily seen to be the excess of the maximum at α over the minimum.

It is clear that the maximum of $f(x)$, at α, is the greatest of the numbers $\overline{f(\alpha+0)}, \overline{f(\alpha-0)}, f(\alpha)$, and that the minimum is the least of the numbers

$\underline{f(\alpha+0)}, \underline{f(\alpha-0)}, f(\alpha)$; and thus that the *saltus* at α is the excess of the greatest over the least of the numbers

$$\overline{f(\alpha+0)}, \underline{f(\alpha+0)}, \overline{f(\alpha-0)}, \underline{f(\alpha-0)}, f(\alpha).$$

At a point of continuity of $f(x)$ the saltus is zero. Any point at which the saltus has a finite value (> 0), or is indefinitely great, is called a *point of discontinuity* of $f(x)$, and in the latter case it is said to be a *point of infinite discontinuity*.

If the closed neighbourhood $(\alpha, \alpha+\delta)$, on the right of α, be taken, the lower limit of the fluctuation in this neighbourhood when δ is indefinitely diminished is called *the saltus at α, on the right*. This is equivalent to the excess of the greatest over the least of the three numbers

$$\overline{f(\alpha+0)}, \underline{f(\alpha+0)}, f(\alpha).$$

A corresponding definition applies to the saltus at α on the left.

225. The points of discontinuity of a function may be classified as follows:—(1) If both the limits $f(\alpha+0)$, $f(\alpha-0)$ exist, and have values which differ from one another, the point α is said to be a point of discontinuity *of the first kind*, or a point of *ordinary* discontinuity.

The difference between the greatest and least of the three numbers $f(\alpha+0), f(\alpha-0), f(\alpha)$ is the saltus, or measure of discontinuity, of the function at α. If α be not a point of the domain of x, $|f(\alpha-0)-f(\alpha-0)|$ measures the saltus at α; and if α be a point of the domain, and $f(\alpha)$ lies between $f(\alpha+0)$ and $f(\alpha-0)$, then the saltus is also measured by $|f(\alpha+0)-f(\alpha-0)|$.

When $f(\alpha)$ does not lie between $f(\alpha+0)$ and $f(\alpha-0)$, the function is said to have an *external saltus* at α.

In every case, the saltus on the right is measured by $|f(\alpha+0)-f(\alpha)|$; and that on the left by $|f(\alpha-0)-f(\alpha)|$.

Whether there be an external saltus at α, or not, the number

$$|f(\alpha+0)-f(\alpha-0)|$$

is said to measure the *oscillation* (Schwingung) at α. The oscillation at a point differs from the saltus in that the functional value $f(\alpha)$, at the point, is in the former case disregarded.

If $f(\alpha)=f(\alpha-0)$, whilst $f(\alpha)\neq f(\alpha+0)$, the function is said to be *ordinarily discontinuous at α on the right*. If $f(\alpha)\neq f(\alpha-0)$, whilst $f(\alpha)=f(\alpha+0)$, the function is said to have an *ordinary discontinuity at α on the left*.

It may happen that $f(\alpha+0), f(\alpha-0)$ have equal values which differ from $f(\alpha)$. In that case the discontinuity at α is *removable* (see § 220); since by merely altering the functional value at the one point α, the function can be made continuous at the point.

(2). If neither of the limits $f(\alpha + 0)$, $f(\alpha - 0)$ exists, the discontinuity at α is said to be *of the second kind.*

The *oscillation** at α is measured by the excess of the greater of the two numbers $\overline{f(\alpha + 0)}$, $\overline{f(\alpha - 0)}$ over the lesser of the two numbers $\underline{f(\alpha + 0)}$, $\underline{f(\alpha - 0)}$, the value of $f(\alpha)$ being left out of account.

The differences $\overline{f(\alpha + 0)} - \underline{f(\alpha + 0)}$, $\overline{f(\alpha - 0)} - \underline{f(\alpha - 0)}$ may be spoken of as the oscillation at α on the right, and on the left, respectively.

By Dini†, a definition of the saltus is adopted which differs from the one which we have employed; he takes the greatest of the four differences $|\overline{\underline{f(\alpha \pm 0)}} - f(\alpha)|$ as the measure of the saltus, the greater of the two differences $|\overline{\underline{f(\alpha + 0)}} - f(\alpha)|$ being taken as the measure of the saltus on the right.

(3). It may happen that one of the two limits $f(\alpha + 0)$, $f(\alpha - 0)$ exists, whilst the other does not. In this case, the point α may be said to be a *point of mixed discontinuity.*

If $f(\alpha)$ exist, and be equal to that one of the two limits $f(\alpha + 0)$, $f(\alpha - 0)$ which exists, then the function is continuous at α on one side, and has a discontinuity of the second kind on the other side.

(4). If one or more of the four limits $\overline{\underline{f(\alpha \pm 0)}}$ be indefinitely great, the point α is one of infinite discontinuity.

Under infinite discontinuities is sometimes included the case in which $f(\alpha)$ is defined by $1/f(\alpha) = 0$, or when $f(x)$ is defined as the limiting sum of a series which, for the value α, becomes divergent.

226. In an arbitrarily small neighbourhood $(\alpha, \alpha + h)$, on the right of a point α at which the limits $\overline{f(\alpha + 0)}$, $\underline{f(\alpha + 0)}$ have different values, there must be an infinite number of points at which $f(x) > \overline{f(\alpha + 0)} - \epsilon$, where ϵ is an arbitrarily small fixed number.

For, if there were only a finite number of such points in $(\alpha, \alpha + h)$, h could be chosen so small that all such points would be excluded from the neighbourhood; thus, in a sufficiently small neighbourhood $(\alpha, \alpha + h)$, we should have at every internal point $f(x) \leq \overline{f(\alpha + 0)} - \epsilon$; and then the upper limit at α on the right could not be $\overline{f(\alpha + 0)}$. In a similar manner it can be shewn that, in the arbitrarily small neighbourhood $(\alpha, \alpha + h)$, there must be an infinite number of points at which $f(x) < \underline{f(\alpha + 0)} + \epsilon$.

---

* This definition of the "Schwingung" is given by Pasch in his *Einleitung in die Differential- und Integralrechnung*, p. 139.

† See *Grundlagen*, p. 55.

In this case we say that, in the arbitrarily small neighbourhood of a, on the right, the function makes an infinite number of finite oscillations. In case of the infinity of one, or of both, of the limits $\overline{f(a-0)}$ and $\underline{f(a+0)}$, and, in the latter case, only if they be of opposite signs, the function makes an infinite number of infinite oscillations in the arbitrarily small neighbourhood of a. A similar remark applies to the case in which $\overline{f(a-0)}$, $\underline{f(a-0)}$ have unequal values. It has thus been shewn that:

A point of discontinuity of the second kind is one such that, in its arbitrarily small neighbourhood, the function makes an infinite number of finite or infinite oscillations.

In an arbitrarily small neighbourhood on either side of a point of discontinuity of the first kind, the function may make an infinite number of oscillations; but since the neighbourhood can be chosen so small that the fluctuation of the function in its interior is arbitrarily small, the oscillations, when they are infinite in number, are arbitrarily small, sufficiently near the point.

EXAMPLES

1. Let $f(x)=\sin x/x$, when $x\neq 0$, and $f(x)=A$, when $x=0$.

In this case $f(+0)=f(-0)=1$, $f(0)=A$; thus $f(x)$ has a removable discontinuity at $x=0$, unless $A=1$, in which case the function is continuous in any interval.

2. Let $f(x)=\dfrac{1}{x-a}$; we have then $f(a+0)=\infty$, $f(a-0)=-\infty$, and $f(a)$ is undefined.

3. Let $f(x)=(x-a)\sin\dfrac{1}{x-a}$, and $f(a)=0$; then $f(a+0)=0$, $f(a-0)=0$. This function is continuous at $x=a$, and makes an infinite number of oscillations in any neighbourhood of that point.

4. If $f(x)=\dfrac{1}{x-a}\operatorname{cosec}\dfrac{1}{x-a}$, then

$$\overline{f(a+0)}=\infty,\quad \underline{f(a+0)}=-\infty,\quad \overline{f(a-0)}=\infty,\quad \underline{f(a-0)}=-\infty.$$

This function has an infinite discontinuity of the second kind at the point a.

5. If $f(x)=e^{\frac{1}{x}}$, we have $f(+0)=\infty$, $f(-0)=0$. If $f(x)=\dfrac{1}{1-e^{\frac{1}{x}}}$, then $f(+0)=0$, $f(-0)=1$.

6. If $f(x)=\sin x$, $\lim\limits_{x\sim\infty} f(x)$ is indeterminate, the limits of indeterminancy being $+1$, -1. In the case $f(x)=x\sin x$, the corresponding limits of indeterminancy are $+\infty$, $-\infty$.

7. Let $y=E(x)$, where $E(x)$ denotes the integral part of x. This function is discontinuous when x has an integral value n; we then have $E(n-0)=n-1$, $E(n)=n$, $E(n+0)=n$.

8. Let (x) denote the positive or negative excess of x over the nearest integer; and when x exceeds an integer by $\frac{1}{2}$, let $(x)=0$. This function is continuous except for values $x=n+\frac{1}{2}$, where n is an integer. We have $(n+\frac{1}{2})=0$, $(n+\frac{1}{2}-0)=\frac{1}{2}$, $(n+\frac{1}{2}+0)=-\frac{1}{2}$.

ORDINARY DISCONTINUITIES

227. Although the points of discontinuity of the first kind have, in accordance with usage, been spoken of above as ordinary discontinuities, it will be shewn that they are, in a sense, more exceptional than discontinuities of the second kind. The following theorem will be established:

The set of points at which a function $f(x)$ has ordinary discontinuities is enumerable (or finite, or absent); whereas the set of points of discontinuity of the second kind may be unenumerable.

Let k be a positive number, and let S_k denote that set of points of ordinary discontinuity in (a, b) at which the saltus of $f(x)$ is $> k$. If the set S_k be not finite, let α be a point of its derivative. Any neighbourhood of α contains points of S_k, and therefore contains pairs of points, both on the same side of α, such that the difference of the values of $f(x)$ at the points of such a pair is numerically $> k$. It follows that the point α must be a point of discontinuity of the second kind, and thus that α is not a point of S_k; and therefore the set S_k is an isolated set, and is consequently enumerable. If $k_1, k_2, \ldots$ be a sequence of diminishing numbers that converges to zero, all the sets $S_{k_1}, S_{k_2}, \ldots$ are enumerable (or finite, or non-existing), and therefore their outer limiting set, which contains all the points at which $f(x)$ has an ordinary discontinuity, is enumerable (or finite, or absent).

THE SYMMETRY OF FUNCTIONAL LIMITS

228. The question as to the nature of the set of points at which a function $f(x)$ has functional limits, on one side, which differ in value from the corresponding functional limits on the other side has been investigated* by W. H. Young.

The general result obtained is that:

Except at points of an enumerable set, the upper and lower limits of a function, on the right, are equal to the upper and lower limits respectively, on the left, and there is no external saltus.

Let us consider the set of points $G_{k,A}$ at which

$$\overline{f(x+0)} - \overline{f(x-0)} > k,\ \overline{f(x+0)} > A;$$

where A and k are positive numbers. If the set $G_{k,A}$ is not enumerable, it contains a component H, each point of which is a limiting point on both sides (§ 89). It follows that, at each point of H, $\overline{f(x-0)} \geq A$; and therefore $\overline{f(x+0)} > A + k$, at each point of H. Similarly, it is seen that, at each point of H, $\overline{f(x+0)} > A + 2k$; and by repeating the process we have $\overline{f(x+0)} > A + nk$, whatever value n may have. It then follows that at the points of H, $\overline{f(x+0)} = +\infty$, and therefore $\overline{f(x-0)} = -\infty$; which is contrary to the condition

---

* See *Quarterly Journal of Math.* vol. XXXIX (1908), p. 67.

$$\overline{f(x+0)} - \overline{f(x-0)} > k.$$

Therefore the set $G_{k\,A}$ is enumerable. Giving to A the values in a sequence of decreasing numbers diverging to $-\infty$, we see that the set of all points at which $\overline{f(x+0)} - \overline{f(x-0)} > k$ is enumerable. Next, letting k have the values of the numbers in a decreasing sequence converging to zero, we see that the set of points at which $\overline{f(x+0)} - \overline{f(x-0)} > 0$ is enumerable. Similarly, it may be shewn that the set of points at which $\overline{f(x-0)} - \overline{f(x+0)} > 0$ is enumerable. Therefore, except at the points of an enumerable set, we have $\overline{f(x+0)} = \overline{f(x-0)}$.

By considering the function $-f(x)$, for which an upper limit is obtained by changing the sign of the corresponding lower limit of $f(x)$, we see that $\underline{f(x+0)} = \underline{f(x-0)}$, except at the points of an enumerable set.

Therefore $\overline{f(x+0)} = \overline{f(x-0)}$ and $\underline{f(x+0)} = \underline{f(x-0)}$, except for values of x belonging to an enumerable set.

In the proof of the above theorem $f(x)$ may be substituted for $\overline{f(x-0)}$, without affecting the reasoning. It then follows that the set of points at which $f(x) > \overline{f(x-0)}$ is an enumerable set. By changing $f(x)$ into $-f(x)$, it then follows that the set at which $f(x) < \underline{f(x-0)}$ is also enumerable. Similarly the sets at which $f(x) > \overline{f(x+0)}$, and at which $f(x) < \underline{f(x-0)}$, are enumerable. Hence, except at the points of an enumerable set, we have

$$\overline{f(x+0)} = \overline{f(x-0)} \geqq f(x)$$

and

$$\underline{f(x+0)} = \underline{f(x-0)} \leqq f(x).$$

The theorem has now been completely established.

Also, it follows that:

The set of points at which there is a removable discontinuity is enumerable (or finite, or absent).

FUNCTIONS CONTINUOUS IN AN OPEN INTERVAL

229. The theorem that a function that is continuous in a closed interval (a, b) attains its upper and lower boundaries (which are also limits), each at one point in the interval, may be extended to the case of a function which is continuous only in an open interval (a, b).

We have in that case the following theorem:

If a function $f(x)$ be continuous in an open interval (a, b), and if $\overline{f(a+0)}$, $\overline{f(b-0)}$ are both less than the upper boundary of the function in the interval, *there is one interior point at least at which the function attains the value of its upper boundary. Similarly if* $\underline{f(a+0)}$, $\underline{f(b-0)}$ *are both greater than the lower boundary of the function in the interval, there is one interior point at least at which the function attains the value of its lower boundary.*

For ϵ, ϵ' may be so chosen that, for $a < x \leq a + \epsilon$, and for $b - \epsilon' \leq x < b$, the upper boundaries of the function are less than in (a, b). The original theorem may then be applied to the closed interval $(a + \epsilon, b - \epsilon')$, and the result follows.

The theorem of § 214 may also be extended to the case of a function that is continuous in an open interval (a, b). We have the following theorem:

If $f(x)$ *be continuous in the open interval* (a, b), *and if any one of the functional limits at* a *is of opposite sign to any one of those at* b, *there is at least one point of the open interval at which* $f(x)$ *has the value zero.*

Let $\overline{f(a+0)} \geqq A \geqq \underline{f(a+0)}$, and $\overline{f(b-0)} \geq B \geqq \underline{f(b-0)}$, where A and B are any functional limits at a and b, not necessarily extreme limits; and suppose $A < 0, B > 0$. Positive numbers ϵ, ϵ' can be so determined that

$$f(a+\epsilon) < 0, f(b-\epsilon') > 0;$$

if the theorem of § 214 be applied to the closed interval $(a + \epsilon, b - \epsilon')$ the result follows.

If A, B *satisfy the conditions*

$$\overline{f(a+0)} \geqq A \geqq \underline{f(a+0)}, \quad \overline{f(b-0)} \geqq B \geqq \underline{f(b-0)},$$

there is at least one value of x *in the open interval at which* $f(x)$ *has a prescribed value* C, *lying between* A *and* B.

For the theorem just proved can be applied to the function $f(x) - C$.

If $f(x)$ *be continuous in the open interval* (a, b), *the aggregate of functional limits, on the right at* a, *contains every number in the interval*

$$(\overline{f(a+0)}, \underline{f(a+0)}).$$

Let c be a number such that $\overline{f(a+0)} > c > \underline{f(a+0)}$, and let ϵ be less than $\overline{f(a+0)} - c$, and than $c - \underline{f(a+0)}$; thus

$$\overline{f(a+0)} - \epsilon > c > \underline{f(a+0)} + \epsilon.$$

In an interval $(a, a + h)$, where $a + h < b$, there is a set of points $G_h^{(1)}$, at each point of which $\overline{f(a+0)} + \epsilon > f(x) > \overline{f(a+0)} - \epsilon$, and another set $G_h^{(2)}$, at each point of which $\underline{f(a+0)} + \epsilon > f(x) > \underline{f(a+0)} - \epsilon$. Let ξ_1 be the upper boundary of the set $G_h^{(1)}$, and ξ_2 the upper boundary of the set $G_h^{(2)}$. We have then, since $f(x)$ is continuous at ξ_1 and ξ_2,

$$\overline{f(a+0)} + \epsilon \geqq f(\xi_1) \geqq \overline{f(a+0)} - \epsilon,$$
$$\underline{f(a+0)} + \epsilon \geqq f(\xi_2) \geqq \underline{f(a+0)} - \epsilon.$$

In the closed interval of which ξ_1, ξ_2 are the end-points, since c lies between $f(\xi_1)$, $f(\xi_2)$, there is a set of points at which the function is equal to c. This set must be closed. if ξ be its upper boundary, we have $f(\xi) = c$; where ξ is a definite point in $(a, a + h)$. If now we assign to h successively the values in a monotone sequence that converges to zero, we obtain a sequence of points ξ that converges to a, and at each point of the sequence the value of $f(x)$ is c. Thus the aggregate of functional limits at a includes the number c. This sequence of values of ξ is determinate, and its construction does not involve the use of the principle of selection.

If we assign to ϵ the values in a monotone sequence that converges to zero, there will be a corresponding sequence of values of ξ_1 such that, in that sequence, $f(\xi_1)$ converges to $\overline{f(a - 0)}$. Similarly the sequence $f(\xi_2)$ will converge to $\underline{f(a + 0)}$.

SEMI-CONTINUOUS FUNCTIONS

230. The condition of continuity of a function $f(x)$, at a point x, namely that, if ϵ be any prescribed positive number, an open interval $(x - h, x + h)$ exists such that, for any point x' in it, $|f(x') - f(x)| < \epsilon$, can be divided into two separate conditions, viz. that $f(x') < f(x) + \epsilon$, and that

$$f(x') > f(x) - \epsilon.$$

It is possible that, at a point x, one of these conditions may be satisfied and not the other. This consideration gives rise to the definition of a property called semi-continuity.

If $\phi(x)$ be a function defined for a continuous domain, and if, corresponding to every arbitrarily chosen positive number ϵ, an open neighbourhood $(x - h, x + h)$ of a particular point (x) can be determined such that, for every point x' in this open interval, the condition,

$$\phi(x') < \phi(x) + \epsilon,$$

is satisfied; then the point x is said to be a point* of *upper semi-continuity* of the function $\phi(x)$.

If an open neighbourhood of the point x can be determined, for each ϵ, such that $\phi(x') > \phi(x) - \epsilon$, then the point x is said to be a point of *lower semi-continuity* of the function $\phi(x)$.

That a point x may be a point of continuity of the function $\phi(x)$, it is necessary that both the above conditions be satisfied.

A function may have upper, or lower, semi-continuity at a point, on the right, or on the left.

---

* See Baire's memoir "Sur les fonctions des variables réelles," *Annali di Mat.* (3 A), vol. III, 1899.

If every point of the domain (a, b), for which the function $\phi(x)$ defined, is a point of upper semi-continuity, then the function $\phi(x)$ said to be an *upper semi-continuous function.*

A similar definition applies to a *lower semi-continuous function.*

It is clear that, if $\phi(x)$ be a lower semi-continuous function, th $-\phi(x)$ is an upper semi-continuous function. Thus the properties of t one class of functions may easily be extended to the other class.

If $f(x)$ be a function, defined for the interval (a, b), and if $M(x)$, m (denote the maximum and the minimum of $f(x)$ at the point x, then $M(x)$ an upper semi-continuous function, and $m(x)$ is a lower semi-continuo function.

For a neighbourhood $(x - h, x + h)$ of any point x can be determine such that the value (see § 224) of $f(x)$, at every point in this neighbou hood, is less than $M(x) + \epsilon$, where ϵ is a prescribed positive number. A every point in $(x - h_1, x + h_1)$, where h_1 is chosen $< h$, the value of th function $M(x)$ is less than $M(x) + \epsilon$. Since this holds for every valu of ϵ, the function $M(x)$ is upper semi-continuous at x.

It is clear that the function $m(x)$, where $m(x)$ denotes the minimum of $f(x)$ at the point x, is a lower semi-continuous function.

The saltus $M(x) - m(x)$, of the function $f(x)$, may be taken to be the value of a function, $\omega(x)$, which is called the *saltus-function* of $f(x)$.

The saltus-function $\omega(x)$ of any function $f(x)$ is an upper semi-continuous function.

For $M(x)$ and $-m(x)$ are both upper semi-continuous functions, and it is easily seen that the sum of two such functions belongs to the same class.

Instead of considering the functions $M(x)$, $m(x)$ we may consider the upper and lower associated functions $A(x)$, $a(x)$ (see § 220). It can be proved in the same manner as in the cases of $M(x)$, $m(x)$ that the upper associated function $A(x)$ is upper semi-continuous, and that the lower associated function $a(x)$ is lower semi-continuous.

The function $\overline{f(x + 0)}$ is upper semi-continuous on the right, and $\underline{f(x + 0)}$ is lower semi-continuous on the right.

If $\phi(x)$ be any upper semi-continuous function, then the set of points for which $\phi(x) \geqq a$ is a closed set, where a is any fixed number.

For let S be a set of points at which the condition $\phi(x) \geqq a$ is satisfied, and let P be a limiting point of S. Let us suppose that, if possible, $\phi(P) < a$; then a neighbourhood of P can be determined, such that at every point in it the value of $\phi(x)$ is less than a, and hence this neighbourhood cannot contain any of the points of S; but this is contrary to the hypothesis that P is a limiting point of S. It follows that $\phi(P) \geqq a$, and thus that the set of points for which $\phi(x) \geqq \alpha$ contains all its limiting points, and is therefore a closed set.

It can be shewn, in a similar manner, or it can be deduced from the above theorem, that *the set of points for which $\psi(x) \leqq \alpha$, is a closed set; where $\psi(x)$ is any lower semi-continuous function.*

If we apply the theorem proved above to the saltus-function $\omega(x)$, of any function $f(x)$, we obtain the following theorem:

Having given any function $f(x)$ defined for a continuous domain, the saltus-function $\omega(x)$ is such that the set of points for which $\omega(x) \geqq \alpha$, forms a closed set.

231. *In an open interval, $f(x)$ and the minimal function $m(x)$ have the same lower boundary.*

For, if there are points at which $f(x)$ is between $A - \epsilon$, $A + \epsilon$: at any such point $m(x) \leqq A + \epsilon$. If A is the lower boundary of $f(x)$, we see that the lower boundary of $m(x)$ is $\leqq A$, since ϵ is arbitrarily small. Again if there are points at which $m(x)$ is between $A' - \epsilon$, $A' + \epsilon$, there must be a point at which $f(x)$ is also between these numbers: hence if A' is the lower boundary of $m(x)$, the lower boundary of $f(x)$ is $< A' - \epsilon$, and is therefore $\leqq A'$. The two inequalities shew that $f(x)$ and $m(x)$ have the same lower boundaries.

If $\phi(x)$ is an upper semi-continuous function, defined in the interval (a, b), the set of points at which $\omega(x) \geqq \sigma$ is non-dense in the interval (a, b).

We have in this case $M(x) = \phi(x)$, and thus $\omega(x) = \phi(x) - m(x)$. If possible, let there be an open interval in which at every point

$$\phi(x) - m(x) \geqq \sigma;$$

and thus $m(x) \leqq \phi(x) - \sigma$. If α be a point of the interval, and $\epsilon\ (< \sigma)$ be arbitrarily chosen, a point x_1 in an arbitrarily small neighbourhood of α, contained in the interval, exists such that

$$\phi(x_1) < m(\alpha) + \epsilon < \phi(\alpha) - (\sigma - \epsilon).$$

In an arbitrarily small neighbourhood of x_1, also contained in the open interval, a point x_2 can be determined such that $\phi(x_2) < \phi(x_1) - (\sigma - \epsilon)$, or that

$$\phi(x_2) < \phi(\alpha) - 2(\sigma - \epsilon).$$

Proceeding in this manner, a point x_n, arbitrarily near to α, can be so determined that $\phi(x_n) < \phi(\alpha) - n(\sigma - \epsilon)$, for any integer n; and thus $\phi(x_n)$ is an arbitrarily great negative number. It follows that there exists an everywhere dense set of points at which $\phi(x) < -A$, where A is an assigned positive number. For, if β be a point at which $\phi(x) < -A'$, where $A' > A$, a neighbourhood of β exists in which

$$\phi(x) < -A' + \epsilon < -A,$$

if ϵ be chosen $< A' - A$. This neighbourhood contains no points of the set for which $\phi(x) \geqq -A$. Since the points β are everywhere dense, we see that, in any interval, a sub-interval exists which contains no points at which $\phi(x) \geqq -A$, and thus the set of such points is non-dense. Giving to A the values in a divergent sequence of increasing numbers, and remembering that $\phi(x)$ is finite at every point, we see that the points of the continuum (a, b) are divided into the sum of sets of a sequence of non-dense sets. But this is impossible, since the continuum does not form a set of the first category. It follows that there cannot exist an everywhere dense set of points, corresponding to each positive number A, such that $\phi(x) < -A$. Thus no open interval exists in which $\omega(x) \geqq \sigma$, and therefore the set of points at which $\omega(x) \geqq \sigma$ is non-dense. We have now the equivalent form of the above theorem:

The points of discontinuity of an upper semi-continuous function form a set of the first category.

232. *If $\phi(x)$ be an upper semi-continuous function, and if, at every point of the interval (a, b), the minimum of $\phi(x)$ be zero, then there exists a set of points, everywhere dense in the interval, at which $\phi(x)$ is itself zero.*

To the interval (a, b), the field of the variable x, let there be fitted on a system of nets with closed meshes. Let P be the centre of a mesh d_n, of D_n; then since the minimum of $\phi(x)$ at P is zero, a mesh d_{n_1} $(n_1 > n)$, of the sequence of meshes which defines P, must be such that in every point of d_{n_1} we have $0 \leqq \phi(x) < \epsilon/2$. Another mesh d_{n_2}, contained in d_{n_1}, can be so determined that, for all points of it, $0 \leqq \phi(x) < \epsilon/2^2$; and so on. We have then a sequence of meshes $d_{n_1}, d_{n_2}, \ldots$ each of which contains the next, and such that, in d_{n_r}, we have $0 \leqq \phi(x) < \epsilon/2^r$. This sequence defines a point $P(\bar{x})$, in all of them, for which $0 \leqq \phi(\bar{x}) < \epsilon/2^r$, for every value of r. Therefore $\phi(\bar{x}) = 0$; and every mesh d_n contains a point such as $\bar{x}$. It is thus seen that $\phi(x)$ vanishes at the points of an everywhere dense set.

In particular, we see that, *if $\omega(x)$ be the saltus-function of any given function $f(x)$, and if $\omega(x)$ has its minimum equal to zero at every point of the domain of x, then $\omega(x)$ vanishes at an everywhere dense set of points. The points of this set are the points of continuity of $f(x)$.*

233. If infinite values of a function are admitted, the function $f(x)$ may still be regarded as semi-continuous, when it is infinite.

Consider a function $f(x)$ which has infinite values; the function

$$\phi(x) = \frac{f(x)}{1 + |f(x)|}$$

is bounded, and the values $1, -1$ correspond to values $+\infty, -\infty$ respectively, of $f(x)$.

We have
$$f(x') - f(x) = \frac{\phi(x') - \phi(x)}{\{1 - |\phi(x')|\}\{1 - |\phi(x)|\}},$$
provided $\phi(x)$, $\phi(x')$ have the same sign; hence if $\phi(x') - \phi(x) < \epsilon$, then
$$f(x') - f(x) < \frac{\epsilon}{\{1 - |\phi(x)|\}^2} < \epsilon'.$$
Therefore, if $\phi(x)$ is upper semi-continuous at x, so also is $f(x)$. In case $\phi(x)$ is positive, and $\phi(x')$ negative, $\phi(x') < \phi(x)$, and then $f(x') < f(x)$.

If $f(x) = \infty$, we have $\phi(x) = 1$, and at all other points $\phi(x') \leq \phi(x)$; therefore a point where $f(x) = +\infty$ is a point where $\phi(x)$ is upper semi-continuous; and similarly where $f(x) = -\infty$, $\phi(x)$ is lower semi-continuous.

It is therefore convenient to regard $f(x)$ as upper semi-continuous, where it has the value $+\infty$, and lower semi-continuous where it is $-\infty$; the correspondence with $\phi(x)$ being thus preserved.

234. The following theorem is a generalization* of the theorem of § 217, that a continuous function is uniformly continuous in its closed domain.

If $f(x)$ is bounded in the interval (a, b), and if k be a number greater than the upper boundary of $\omega(x)$ in (a, b), there exists a positive number α, such that, in every closed sub-interval in (a, b) of length not exceeding α, the fluctuation of $f(x)$ is $< k$.

It is convenient to assume that, for values of x that are $< a$, $f(x)$ is defined to be equal to $f(a)$, and that, for values of x that are $> b$, $f(x) = f(b)$.

If x be any point in (a, b), the fluctuation of $f(x)$ in the interval $(x - h, x + h)$ is $< k$, provided h be sufficiently small. The upper boundary of all values of h for which this is the case may be denoted by $\bar{h}$. If θ be a fixed number such that $0 < \theta < 1$, the fluctuation of the function in the closed interval $(x - \theta\bar{h}, x + \theta\bar{h})$ is $< k$. Corresponding to each point x in (a, b), a definite interval may be assigned, viz. the interval $(x - \theta\bar{h}, x + \theta\bar{h})$, or the part of it which is interior to (a, b). By the Heine-Borel theorem, a finite set of these intervals exists, such that every point of (a, b) is interior to at least one of these intervals. The end-points of the intervals of this finite set Δ form a finite set of points in (a, b); let α be the smallest of the distances between consecutive points of this set. Any closed interval whatever, in (a, b), of length not exceeding α is in one of the intervals of the finite set Δ. Therefore, in such an interval, the fluctuation of $f(x)$ is $< k$.

The corresponding theorem for the case of a bounded function defined in a cell of two or more dimensions may be stated as follows:

If a function be bounded in a cell (a, b) of any number of dimensions, and if k be a number greater than the upper boundary of the saltus-function in the

---

* See Baire, *loc. cit.* p. 15.

whole cell, there exists a positive number α such that, in every closed cell contained in (a, b), of span not exceeding α, the fluctuation of the function is $< k$.

The proof is a modification of that given above for the case of a linear interval. There is shewn to exist a finite set of closed cells, such that in each of them the fluctuation of the function is $< k$, and such that each point in the cell (a, b) is interior to one of the cells of the finite set. If we project on the edges of (a, b) all the edges of the cells of this finite set, we have on each edge of (a, b) a finite set of points. The number α can then be taken to be the least distance between consecutive points of these finite sets, when all the points on all the edges of (a, b) are taken into account. Any cell whatever, in (a, b), of span not exceeding α will be in one of the cells of the finite set, and therefore the fluctuation of the function in such a cell is $< k$.

The definitions of semi-continuity given above, and the theorems established, are applicable when the domain of the variable x is not an interval (a, b), but any closed set of points. In that case, we regard functional values in any interval as only existing at those points in the interval which belong to the domain of x.

Moreover, the definitions and theorems are applicable to the case of functions of a number p of variables. In this case, instead of an interval $(x - h, x + h)$ used in defining semi-continuous functions and the saltus-function, we may employ the "sphere"

$$(\xi^{(1)} - x^{(1)})^2 + (\xi^{(2)} - x^{(2)})^2 + \ldots + (\xi^{(p)} - x^{(p)})^2 < h^2, \text{ or } \leq h^2,$$

according as the neighbourhood is open or closed, or else a cell may be employed.

EXAMPLE

If* $f(x)$ be any function, $\phi(x)$ its maximum, and $\psi(x)$ its minimum at the point x, and $g(x)$ be any continuous function, then $\phi(x) + g(x)$, $\psi(x) + g(x)$ are the maxima and minima at x, of $f(x) + g(x)$. Also the functions $\phi(x) - f(x)$, $f(x) - \psi(x)$ have each, in any domain, the minimum zero.

APPROXIMATE CONTINUITY

235. A function $f(x)$ is said† to be *approximately continuous* at a point α of the domain (a, b) of the variable x, when, corresponding to each arbitrarily chosen number ϵ, the set of points x for which $| f(x) - f(\alpha) | < \epsilon$ has the metric density unity at the point α.

The necessary and sufficient condition that the function $f(x)$ should be approximately continuous at the point α is that it should be continuous, at α, relatively to a set of points G which has the metric density unity at the point α.

---

* Baire, *loc. cit.* p. 9.

† Denjoy, *Bull. de la soc. math. de France*, vol. XLIII (1915), p. 165. See also Looman, *Fundamenta Math.* vol. V (1924), p. 105.

The condition in the theorem is sufficient; for there then exists a number h, such that, in the interval $(a - h, a + h)$, the condition

$$|f(x) - f(a)| < \epsilon$$

is satisfied for every point x that belongs to G. This part of G has the same metric density at a as G itself.

To shew that the condition is necessary, let it be assumed that the set $G(a, \epsilon)$ of points at which $|f(x) - f(a)| < \epsilon$ has the metric density 1 at the point a, whatever value ϵ may have.

Let $\{\epsilon_n\}$ be a sequence of decreasing values of ϵ that converges to zero: for each ϵ_n a number h_n $(< h_{n-1})$ can be so determined that the measure of the part of $G(a, \epsilon_n)$ in the interval $(a - h_n, a + h_n)$ is $> 2h_n(1 - \epsilon_n)$. The numbers h_n may, if necessary, be so altered that $h_{n-1}/h_n < 1/\epsilon_n$, for every value of n.

Let F_n be the part of $G(a, \epsilon_n)$ that is in the two intervals

$$(a + \epsilon_n h_{n+1}, a + h_n), (a - h_n, a - \epsilon_n h_{n-1});$$

then F_{n+m} is a part of $G(a, \epsilon_{n+m})$, and therefore of $G(a, \epsilon_n)$. If G be the set $M(F_1, F_2, \ldots F_n, \ldots)$, then the set G has the metric density 1 at the point a, since $m(F_n) > 2h_n(1 - 2\epsilon_n)$. Let $x_1, x_2, \ldots x_m, \ldots$ be a sequence of numbers converging (increasing, or decreasing) to a, and such that all of them are points of G. Each one of them x_m belongs to one of the sets $F_1, F_2, \ldots F_n, \ldots$; say to $F_{\overline{m}}$. It is clear that the numbers $\overline{m}$ increase as m does so, and that $\overline{m} \sim \infty$ as $m \sim \infty$.

We have then $|f(x_m) - f(a)| < \epsilon_{\overline{m}}$; and therefore the sequence $\{f(x_m)\}$ converges to $f(a)$. Hence $f(x)$ is continuous relatively to G, also G has the metric density 1 at the point a.

THE CLASSIFICATION OF DISCONTINUOUS FUNCTIONS

236. Let us suppose a function to be defined for all points in a continuous interval (a, b); at each point x, the saltus of the function has a finite value, or is indefinitely great, its value being zero at a point of continuity. With a view to the classification of functions, in accordance with the distribution of the points of continuity, and of discontinuity, in the interval (a, b), the question arises, what is the most general distribution of the points of continuity?

The answer to this question is contained in the theorem:

The points of continuity, if they exist, of a function defined for a continuous interval, form an ordinary inner limiting set.

To prove this theorem, let ϵ be a fixed positive number, and enclose each point P, of continuity of a function $f(x)$, in an interval so chosen that the fluctuation of $f(x)$ therein is $< \epsilon$; all the points of continuity are then enclosed in a set of intervals which in general overlap; let these intervals be replaced by the equivalent set Δ_ϵ, of non-overlapping intervals. Imagine these sets Δ_ϵ constructed, corresponding to a sequence of diminishing values of ϵ which converges to zero; there exists then a set of points which are interior to intervals of all these sets of intervals, since this set of points includes all the points of continuity of $f(x)$. If Q be any point which belongs to the inner limiting set so defined, Q must be a point of continuity of $f(x)$; for, corresponding to any arbitrarily small number ϵ_n, Q is in the interior of some interval in which the fluctuation of the function is $< \epsilon_n$, and thus Q is a point of continuity of the function.

In accordance with the theorems which have been obtained in § 97, relating to ordinary inner limiting sets, the points of continuity of a function may form an enumerable set which contains no component dense in itself, or else they form a set of the cardinal number of the continuum. In the latter case the set is a residual set, provided it be everywhere dense. For the closed set at which $\omega(x) \geq \epsilon$, will, for every value of ϵ, be non-dense; and the points of discontinuity accordingly form a set of the first category.

These results lead to the following classification* of functions:

(1). A function may have no points of continuity; it is then said to be totally discontinuous.

(2). The points of continuity may form an enumerable set which has no component dense in itself.

(3). The set of points of continuity may be of the cardinal number of the continuum, and be

(*a*), non-dense;

(*b*), everywhere dense and unclosed, in which case the function is said to be a *point-wise discontinuous function*;

(*c*), everywhere dense and closed, in which case the function is continuous;

(*d*), everywhere dense in each interval of a set, and non-dense in each interval of another set external to the former one.

This last case (*d*) is not essentially distinct from the former ones.

By Hankel and others the term "totally discontinuous" has been applied to all functions which are neither continuous nor point-wise discontinuous.

237. It has been shewn by W. H. Young that a function can be constructed which is continuous at every point of any given ordinary inner limiting set of points, and is discontinuous at every other point of the interval.

---

* See a paper by W. H. Young. *Wiener Sitzungsber.* vol. CXII. Abt. II a (1903), p. 1307.

Let E denote an ordinary inner limiting set, and let the function $f(x)$ be defined as follows:

(1). At every point x, of E, let $f(x) = x$

(2). It has been shewn that a sequence of sets of non-overlapping open intervals can be constructed such that the only points, each of which is in an interval of every set, are the points of E. Let Q be a limiting point of E which does not belong to E, then a number n exists such that Q is in an interval of the $(n-1)^{\text{th}}$ set, but not in one of the intervals of the n^{th} set. Let this interval of the $(n-1)^{\text{th}}$ set be of length d_Q: and at the point Q let $f(x) = x_Q + e^n d_Q$, where e is a fixed positive number less than unity: in the case $n = 1$, we put $d_Q = e$.

(3). If R be a point which does not belong either to E or to its derivative, it must lie between two definite points A, B both of which belong to E or to E', and such that no point of E or of E' lies between A and B. If x_R be a rational number, let $f(x_R) = x_A$ or x_B, according as R is nearer to A or to B, when x_R is irrational, let $f(x_R) = x_R$: and if R be the middle point of AB, let $f(x_R) = x_R$.

It is clear that the function so defined is discontinuous at every internal point of the interval AB, and at the end-point A it is continuous, or discontinuous, on the right, according as A does, or does not, belong to E; a similar result holds for B. It has thus been shewn that the function is discontinuous at every point which does not belong to E.

To shew that, at any point P, of E, the function is continuous, consider those intervals, one of each set in the sequence, which contain the point P: the lengths of these intervals will have a lower limit d, which may be zero. In every interior point of d we have $f(x) = x$: and thus, if P be interior to d, P is a point of continuity of the function. If P be an end-point of d, it is certainly continuous on the side towards the interval: and we have to shew that it is also continuous on the other side. Choose an arbitrarily small number σ, and an arbitrarily large integer m: then a number $n_1 > m$ can be found such that the $n_1{}^{\text{th}}$ interval, and all subsequent intervals of the sequence which contain P, are of length between d and $d + \sigma$. The piece of one of these intervals which is not a portion of d is of length $< \sigma$; suppose that Q is a point in this piece which belongs to E' but not to E: then

$$|f(x_P) - f(x_Q)| = |x_P - x_Q - e^n d_Q| < |\sigma + e^m (d - \sigma)|$$

since $n \geqq n_1 > m$, and $d_Q < d + \sigma$. From this, there follows

$$|f(x_P) - f(x_Q)| < 2\sigma + e^m d < 3\sigma,$$

if m be chosen sufficiently great. If R be an interior point of an interval AB which contains in its interior no point of E', or of E, and if the points A, B be both so near to P that their distances from it are less than σ, we have $|f(x_P) - f(x_R)| < \sigma$, in virtue of the definition of $f(x_R)$; also if only one of the ends A, B be within the interval of length $< d + \sigma$, which has been chosen, then an interval further on in the sequence can always be found such that the middle point of AB is exterior to it, and thus the inequality $|f(x_P) - f(x_R)| < \sigma$ holds as in the former case. It has now been shewn that, for any arbitrarily chosen σ, a neighbourhood of P can be found such that, for all points x in it, $|f(x_P) - f(x)| < 3\sigma$; therefore P is a point of continuity of the function. The case in which $d = 0$ does not require separate treatment.

EXAMPLES

1.* Let G denote a non-dense perfect set of points in the segment $(0, 1)$, such that the end-points of the complementary intervals are rational points. Let $f(x)$ be defined thus:—At every irrational point inside an interval complementary to G, let $f(x) = x$; at every rational point of such interval, let $f(x)$ be equal to the value of x at the middle point of the interval; and at every point external to a complementary interval, let $f(x) = 1$. This function is discontinuous except at the middle points of the intervals complementary to G; thus the set of points of continuity is an enumerable set which contains no component dense in itself.

2.* With the same non-dense perfect set as in Ex. 1, let AB be a complementary interval of G, and M its middle point. At every rational point of AM except M, let $f(x) = x_A$, and at every rational point of MB except M, let $f(x) = x_B$; also at all points of $(0, 1)$, except those for which the functional value has been already specified, let $f(x) = x$. In this case the points of continuity are non-dense, and of the power of the continuum.

POINT-WISE DISCONTINUOUS FUNCTIONS

238. A function being defined for the continuous domain (a, b), it has been shewn in § 230 that, *if k be any fixed positive number, those points, at which the saltus of the function is $\geqq k$, form a closed set.*

This theorem follows immediately from the property of semi-continuous functions established in § 230, by considering the saltus-function. It may also be proved directly as follows:

If P be a limiting point of the set for which the saltus is $\geqq k$, then in any arbitrarily small neighbourhood of P there are points of the set; hence the fluctuation of the function in this neighbourhood is $\geqq k$, and therefore the saltus at P is $\geqq k$.

Moreover, such a limiting point P, of the set of points at which the saltus is $\geqq k$, must be a point of discontinuity of the second kind, at least on one side of P. If, at P, the function have a limit on the right, a neighbourhood PQ can be found such that the inner fluctuation in PQ is $< k$; hence inside PQ there can be no point at which the saltus is $\geqq k$; and therefore P is not a limiting point on the right, of the set for which the saltus is $\geqq k$. A similar remark applies to the left of P.

---

* See W. H. Young, *loc. cit.*

We have already, in § 236, defined a point-wise discontinuous function as one of which the points of continuity are everywhere dense and unclosed in the domain of the function; this definition is that given by Dini*, and is equivalent to the following definition given by Hankel†:

A point-wise discontinuous function is one for which those points at which the saltus is $\geq k$, *an arbitrarily chosen positive number, form a non-dense set* K, *whatever value* k *may have.*

That this set is closed has been shewn above.

To prove the equivalence of the two definitions, let it be assumed that in any arbitrarily chosen sub-interval (α, β), a point of continuity x_1 can be found. A neighbourhood can be found for x_1, internal to (α, β), in which the fluctuation of the function is $< k$, and this neighbourhood can contain no point at which the saltus is $\geq k$; hence the points at which the saltus is $\geq k$ form a non-dense set K, since, interior to any sub-interval, a sub-interval can be found which contains no point of the set K.

Conversely, choose a descending sequence of values of k, say

$$k_1, k_2, k_3, \ldots$$

which converges to zero, and let $K_1, K_2, K_3, \ldots$ be the corresponding non-dense closed sets, each of which necessarily contains the preceding one; then the set $M(K_1, K_2, K_3, \ldots)$ is the set of all the discontinuities of the function.

In accordance with § 93, this set is of the first category, and the complementary set, which is the set of points of continuity of the function, is everywhere dense, and has the cardinal number of the continuum, being a set of the second category, of the kind that is called a residual set.

It will be observed that the set of all the points of discontinuity may be either everywhere dense, or non-dense, in the whole or part of the domain of the variable. This set may be finite, enumerably infinite, or of the power of the continuum.

In accordance with the theorem proved in § 231, an upper semi-continuous function can only be point-wise discontinuous.

The set K, although non-dense, is not necessarily of content zero. By Harnack‡, the term "point-wise discontinuous function" was only used for such functions as possess the property that the set K, for each value of k, has content zero. It will be seen that this latter case is of special importance in connection with the theory of Riemann integration.

---

* See *Grundlagen*, p. 81.

† *Math. Annalen*, vol. XX (1882), p. 90. This is a reproduction of Hankel's *Univ. Programm*, Tübingen, 1870, entitled "Untersuchungen über die unendlich oft oscillierenden und unstetigen Functionen."

‡ *Math. Annalen*, vol. XIX (1882), p. 242, and vol. XXIV (1884), p. 218.

It has already been shewn, in § 236, that the points of continuity of the point-wise discontinuous function form an ordinary inner limiting set; and if

$$\{\delta_1\}, \{\delta_2\}, \ldots \{\delta_n\}, \ldots$$

be the sets of intervals complementary to the closed sets

$$K_1, K_2, \ldots K_n, \ldots,$$

they form a sequence of sets of non-overlapping intervals which define the set of points of continuity as their inner limiting set.

The whole theory of point-wise discontinuous functions is applicable to the case in which the domain of the variable is not a continuum, but is any perfect set. In this case also, the points of continuity of a point-wise discontinuous function are everywhere dense relatively to the perfect domain, and the points at which the measure of discontinuity is $\geq k$ form a closed set, non-dense relatively to the domain of the variable. That this is the case may be shewn by making the points of the perfect set correspond in order to the points of a continuous interval, as explained in § 96. The points of discontinuity, and those of continuity, relatively to the perfect domain, are sets of the first and the second category respectively, relative to that domain.

239. An important class of functions is that in which each function has ordinary discontinuities only; the domain of the functions being either a continuous interval, or a perfect set of points.

A function which has only ordinary discontinuities is point-wise discontinuous.

The set of points at which the saltus is $\geq k$ must be finite. For if it were not finite it would have a limiting point which would be a point of discontinuity of the second kind; and this does not exist. It thus follows that the set of points of discontinuity is a non-dense enumerable set.

A monotone function is one such that, for every pair of values x_1, x_2 of the variable, such that $x_2 > x_1$, the condition $f(x_2) \geq f(x_1)$ is satisfied; or else, for every such pair the condition $f(x_2) \leq f(x_1)$ is satisfied.

In the former case the function is said to be monotone non-diminishing, and in the latter case monotone non-increasing. Since a monotone function has no oscillations in the neighbourhood of any point, every discontinuity is an ordinary one. It follows that:

A monotone function is point-wise discontinuous (*or continuous*). *Its points of discontinuity form an enumerable set. The points at which the saltus is $\geq k$ (> 0) form a finite set.*

The domain of the variable being either a continuous interval or any perfect set, let us suppose that, at every point, the oscillation (see § 225) of the function, both on the right, and on the left, is $< k$; there can then only be a finite number of points at which the saltus is $\geqq k$; *i.e.* the set K is finite.

For if K were not finite, it must contain a limiting point P, which has been shewn, in § 238, to be a point of discontinuity of the second kind. Any arbitrarily small neighbourhood of P, on one side at least, must therefore contain points at which the saltus is $\geqq k$, and hence the oscillation at P on this side could not be $< k$.

The domain can therefore be divided into a finite number of parts within each of which there is no point at which the saltus is $\geqq k$; and it follows that the domain can be divided into a finite number of parts, within each of which the fluctuation of the function is $< k$.

If, at each point of a set which is everywhere dense in the domain of the variable, there exist a limit of the function on one side at least, then the function is either point-wise discontinuous, or else it is continuous.

In any interval, containing points of the domain, a point can be found which has a neighbourhood, on one side at least, in which the inner fluctuation is $< k$; within such neighbourhood the saltus is everywhere $< k$; hence the points of K are non-dense in the domain, and thus the function is either point-wise discontinuous, or else it is continuous.

If a function be defined for a continuous interval, and all the points of discontinuity be ordinary ones, at least on one side, then the set K, of points at which the saltus is $\geqq k$, is a set of content zero.

The set K can be resolved into a perfect set G and an enumerable set: the set G contains points which are limiting points on both sides, and at such points the oscillation both on the right and on the left must be $\geqq k$: it follows that the set G is non-existent, and that K is therefore an enumerable closed set, which has necessarily content zero.

The theorem still holds if there be points of discontinuity of the second kind which form a closed set of content zero, for these points may be enclosed in a finite number of intervals whose sum is arbitrarily small: the theorem can then be applied to each of the remaining intervals of the domain.

240. It has been shewn in § 222 that the aggregate of functional limits of a function, at any point, is a closed set.

A point-wise discontinuous function can be so constructed that, at a point x_0, the aggregate of functional limits is a prescribed closed set G.*

To establish this theorem, we observe that, if G be unenumerable, it may be replaced by an enumerable set G_1, everywhere dense in G. Let

$$V_1, V_2, V_3, \ldots$$

denote the set G_1, arranged in the order-type ω.

* This theorem was given by Bettazzi, see *Rendiconti di Palermo*, vol. VI (1892), p. 173.

Next, choose an enumerable sequence $x_1, x_2, \dots x_n, \dots$ of values of x having the single limiting point x_0; and arrange the sequence $\{x_n\}$ in the order-type ω^2. The sequence $\{x_n\}$ may thus be split up into an enumerable set of sequences $\{x_{1r}\}, \{x_{2r}\}, \dots \{x_{sr}\}, \dots$ each of which has the limit x_0. The function $f(x)$ may be defined by the specifications, that $f(x) = 0$, for all values of x which do not belong to the sets $\{x_{1r}\}, \{x_{2r}\}, \dots$; and that $f(x_{sr}) = v_{sr}$, where $v_{s1}, v_{s2}, \dots v_{sr}, \dots$ form a sequence chosen so as to converge to the limit V_s. The function $f(x)$ so defined is continuous at every point except $x_0, x_1, \dots x_n, \dots$, and it has the required property; since the points of G_1, and therefore of G, are all values of the aggregate of functional limits at the point x_0.

EXAMPLES

1.[*] If $f(x)$, $\phi(x)$ be two point-wise discontinuous functions defined for the same interval, there is an everywhere dense set of points at each of which both functions are continuous. This theorem follows at once from the fact that the points common to two residual sets also form a residual set, and that this holds for every sub-interval contained in the given interval.

2. Let (x) denote the positive or negative excess of x above the integer nearest to it, and if x be half-way between two successive integers, let $(x)=0$. Let a function[†] $f(x)$ be defined for the interval $(0, 1)$ as the limit of

$$\frac{(x)}{1}+\frac{(2x)}{4}+\frac{(3x)}{9}+\dots+\frac{(nx)}{n^2},$$

when n is indefinitely increased. The function $f(x)$ is a point-wise discontinuous function, in which the set K, of points at which the saltus is $\geqq k$, is finite for each positive value of k. It can be proved that, if $x=m/2n$, where m and $2n$ are relative primes, then

$$f\left(\frac{m}{2n}+0\right)=f\left(\frac{m}{2n}\right)-\frac{\pi^2}{16n^2}, \qquad f\left(\frac{m}{2n}-0\right)=f\left(\frac{m}{2n}\right)+\frac{\pi^2}{16n^2}.$$

For values of x not of the above form, $f(x)$ is continuous. The number of points of K is the number of irreducible proper fractions having even denominators $2n$, such that $\pi^2/8n^2 \geqq k$. The set of all the points of discontinuity is everywhere dense in the interval $(0, 1)$.

3. Let[‡] $y=c$, for all rational values of x; and $y=d$, for all irrational values of x. This function is totally discontinuous.

4. Let[§] $f(x)=1$, for all values of x in the interval $(0, 1)$, except $x=\frac{1}{2^n}$, $(n=1, 2, 3, \dots)$, for which $f(x)=0$. At each of the points $\left(\frac{1}{2^n}\right)$ there is a saltus equal to unity. This function is point-wise discontinuous, and the content of K is zero, for every value of k.

5. In[§] the interval $\left(\frac{1}{2}, 1\right)$ of x, let $f(x)=1$; in the interval $\left(\frac{1}{2^2}, \frac{1}{2}\right)$, let $f(x)=\frac{1}{2}$; and in general, in the interval $\left(\frac{1}{2^{n+1}}, \frac{1}{2^n}\right)$, let $f(x)=\frac{1}{2^n}$. In this case the point-wise discontinuous function $f(x)$ is such that the number of points at which the saltus is $\geqq k$ is finite for every value of $k>0$.

---

[*] See Volterra, *Giornale di Mat.* vol. XIX (1881), p. 77.

[†] See Riemann's *Ges. Werke*, p. 242.

[‡] Dirichlet's *Werke*, p. 132.

[§] Hankel, *Math. Annalen*, vol. XX (1882), p. 85.

6. The* points of a continuous interval (0, 1) may be put into correspondence with the points of a non-dense set of points, dense in itself, contained in an interval (a, b), in such a manner that the relative order of two points of the interval (0, 1) is the same as that of the corresponding points in (a, b). Such a correspondence is defined by a point-wise discontinuous monotone function $y = f(x)$.

7.† Let the numbers of the interval (0, 1) be expressed as finite or infinite decimals $x = \cdot a_1 a_2 a_3 \dots a_n \dots$, and let $f(x) = \left(\frac{a_1}{10}\right)^2 + \left(\frac{a_2}{100}\right)^2 + \dots$. The function $f(x)$ is monotone, and is discontinuous for every value of x represented by a finite decimal. The set of points K_k for a given value of k, is finite. The function $f(x)$ defined by $f(x) = \cdot a_1 0 a_2 0 a_3 \dots$ has similar properties.

8.‡ Let the points of the interval (0, 1) be represented by decimals, and consider the set G_0 of those points for which only the digits 0 and 1 occur in the decimal representation, excluding those points for which all the figures are 0, from and after some fixed place. The set G_0 is non-dense in the interval (0, 1), and has the cardinal number c. Any point x_0, of G_0, is represented by $\cdot a_1 a_2 a_3 \dots a_n \dots$, where a_n is 0 or 1. Let ξ be a fixed point $\cdot b_1 b_2 \dots b_n \dots$ of G_0, and let x_ξ denote $x_0 + 2\xi \equiv \cdot c_1 c_2 \dots c_n \dots$; so that $c_n = a_n + 2b_n$. With ξ fixed, let the set of all points x_ξ be denoted by G_ξ; the points of G_ξ are all different from those of G_0, and for two values ξ, ξ' of ξ, the sets G_ξ, $G_{\xi'}$ have no point in common. For the two numbers $c_1 c_2 \dots c_n \dots$, $c_1' c_2' \dots c_n' \dots$ are identical only when $c_n = c_n'$, which is the case only when $a_n = a_n'$, and $b_n = b_n'$. If we read off in the dyad scale the decimal representation of ξ, we obtain, by giving ξ all the values in G_0, every point in (0, 1) except the point 0, and these once only; let the point which, by thus using the dyad scale, corresponds to ξ, be denoted by (ξ). Now let $f(x)$ be defined by the rules $f(x_\xi) = (\xi)$, and $f(x) = 0$ for all other values of x. The point-wise discontinuous function $f(x)$ so defined for the interval (0, 1) is such that, at all the points of the enumerable set G_ξ, the saltus is (ξ); the set of all the points of discontinuity is non-dense in (0, 1); and $f(x)$ is constant, $= 0$, in an everywhere dense set of linear intervals.

9.‡ Let the points x, of the interval (0, 1), be expressed as radix-fractions in the scale of 3. Let G_0 be the set of points for which all the figures of the radix-fraction are 0 and 1, excepting those points for which all the figures are 0, after some fixed place. Let G_n consist of all the points which contain the digit 2 in at most the first n places, but are also such that the nth figure is 2; then G_n is non-dense, and of cardinal number c. There are left only those points for which the radix-fractions contain the digit 2 an infinite number of times; and these points belong to a set H, for which the radix-fractions contain an infinite number of digits other than 2, or to a set G, for each point of which every digit is 2, from and after some fixed one. Each point of G can be represented by a terminating radix-fraction which contains only a finite number of 2's, and can be added to a G_n. Let $G_0, G_1, G_2, \dots$, when so increased, become $\overline{G}_0, \overline{G}_1, \overline{G}_2, \dots$; and take a sequence of decreasing numbers $g_0, g_1, g_2, \dots$. Let the function $f(x)$ be defined by the rules $f(x) = g_n$, if x is a point of $\overline{G}_n$, and $f(x) = 0$ for all points of H. The point-wise discontinuous function $f(x)$ is continuous at all the points of H, and the points of discontinuity are everywhere dense in (0, 1), and of cardinal number c.

---

* Harnack, *Math. Annalen*, vol. XXIII (1884), p. 285.

† Peano, *Riv. di Mat.* vol. I.

‡ Schoenflies, *Göttinger Nachrichten*, 1899.

DEFINITION OF POINT-WISE DISCONTINUOUS FUNCTIONS BY EXTENSION

241. Let us suppose a function $f(x)$ to be defined for a domain which consists of a set of points which is dense in itself, but not closed; and further let us assume that $f(x)$ is continuous in this domain. The new domain obtained by adding to the original domain those of its limiting points which do not belong to it may be spoken of as the *extended domain.* It has been pointed out in § 222 that, at a point a, of the extended domain, which does not belong to the original domain, there is an aggregate of functional limits which is certainly a closed set, and may consist of a finite, or an infinite, set of numbers.

Let us now define a function $\phi(x)$, for the extended domain, in the following manner.—At each point of the original domain, which may be called a *primary point,* let $\phi(x) = f(x)$; at each point a, which may be called a *secondary point,* and which does not belong to the original domain, attribute to $\phi(x)$ the values contained in the aggregate of functional limits of $f(x)$ at a; this function $\phi(x)$ may then be multiple-valued at any secondary point. The new function $\phi(x)$, defined for the extended domain, may be spoken of as *the function obtained by extension of* $f(x)$; and those points for which $\phi(x)$ is multiple-valued are regarded as points of discontinuity at which the measure of discontinuity is the excess of the greatest over the least value of the function at that point.

It will be shewn that *the extended function* $\phi(x)$ *is point-wise discontinuous in the extended domain, unless it be continuous.* This gives rise to a method of constructing point-wise discontinuous functions which has been employed by Brodén in various special cases. Since we may so choose the original domain that it shall consist of an enumerable set of points, the method includes one for the construction of a point-wise discontinuous function from an enumerable set of specifications.

To prove that the extended function $\phi(x)$ is at most point-wise discontinuous, it is sufficient to shew that $\phi(x)$ is continuous at all points of the original domain G, which is a set that is everywhere dense in the extended domain G'.

Let x be a point of G, and let it be the limiting point of a convergent sequence $(x_1', x_2', x_3', \ldots)$, of which all the points belong to G'. Consider the aggregate $\{\phi(x_1'), \phi(x_2'), \ldots\}$, where $\phi(x_1'), \phi(x_2'), \ldots$ have any of the values which belong to the points $x_1', x_2', x_3', \ldots$. Now a point x_n, of G, can be found such that $|x_n' - x_n| < \eta_n$, and $|\phi(x_n') - f(x_n)| < \epsilon_n$, where η_n, ϵ_n are independent arbitrarily small numbers. If we take a sequence of values of η, such that $\eta_1 > \eta_2 > \eta_3 > \ldots$, with zero as its limit, and also a similar sequence of the ϵ numbers, then the sequence $(x_1, x_2, x_3, \ldots)$ has the same limit x as the sequence $(x_1', x_2', x_3', \ldots)$; and the aggregate

$\{\phi(x_1'), \phi(x_2'), \ldots\}$ has the same limit as the convergent aggregate $\{f(x_1), f(x_2), \ldots\}$, viz. $f(x)$ or $\phi(x)$; and thus the theorem is established.

It will be observed that the values of $\phi(x)$ at all the secondary points in an arbitrarily small neighbourhood of a secondary point a depend only on the values of $f(x)$ in that same neighbourhood: it follows therefore that a is a point of continuity or of discontinuity of $\phi(x)$, according as the aggregate of functional limits of $f(x)$ at a consists of one number or of more. In the latter case the measure of discontinuity of $\phi(x)$ at a is the excess of the greatest over the least of the numbers belonging to the values of $\phi(x)$ at the point.

It has been shewn (§ 240) that a point-wise discontinuous function can be so constructed that, at a given secondary point, the values of the function may be an arbitrarily assigned closed set

242. Although a class of point-wise discontinuous functions may be obtained by extension of a continuous function defined for a primary domain, dense in itself but unclosed, yet not every point-wise discontinuous function can be generated in this manner.

Let $f(x)$ be a point-wise discontinuous function in a domain which is either a continuum or a perfect set of points.

Consider the function $\phi(x)$, obtained by taking the values of $f(x)$ as given only at its points of continuity, and extending this function to the complete domain, in the manner explained above.

At each point of discontinuity of $f(x)$ there is a saltus k_f, and at that point the function $\phi(x)$, obtained by extending the set of values of $f(x)$ at its point of continuity, has a measure of discontinuity k_ϕ, which will be zero in case $\phi(x)$ be continuous at the point, but in any case the condition $k_\phi \leqq k_f$ is satisfied, since, within any neighbourhood of the point, the fluctuation of $\phi(x)$ cannot be greater than that of $f(x)$.

If $k_\phi = 0$ at any point of discontinuity of $f(x)$, that point may be said to be a point of *unessential* discontinuity of the function $f(x)$, and if $k_\phi > 0$, the point is one of *essential* discontinuity.

Let now a function $\chi(x)$ be defined for the whole domain as follows:—At every point of continuity of $f(x)$, and at every point of discontinuity at which $k_f = k_\phi$, let $\chi(x) = 0$; at each point at which $k_f > k_\phi$, let

$$\chi(x) = k_f - k_\phi.$$

The function $\chi(x)$ is not necessarily continuous at every point at which it is zero. At a point x_1 at which $\phi(x)$ is continuous, the measure of discontinuity of $\chi(x)$ is k_f, or $\chi(x_1)$; but this is not necessarily the case if $\phi(x)$ be not continuous at x_1. This function $\chi(x)$ may be called a ***point-wise discontinuous null-function***.

By subtracting from $f(x)$ a function $\psi_1(x)$ which never exceeds, at any point x, in absolute value, the value of $\chi(x)$, we obtain a function $\phi_1(x)$, of which the measure of discontinuity is everywhere $= k_\phi(x)$.

The function* $\phi_1(x)$ may be spoken of as the *most nearly continuous function associated with* $f(x)$.

It thus appears that a point-wise discontinuous function can always be expressed as the sum of a point-wise discontinuous null-function and the most nearly continuous function associated with the given function.

The latter function $\phi_1(x)$ has only those discontinuities which necessarily arise from the values of the given function at its points of continuity, and is independent of the parts of the discontinuities which arise out of the functional values of $f(x)$ at the points of discontinuity. The null-function depends upon the unessential parts of the discontinuity of $f(x)$.

EXAMPLES

1.† Let $f(x) = 0$, for $x = 0, \frac{1}{\pi}, \frac{1}{2\pi}, \frac{1}{3\pi}, \ldots$; and for all other positive and negative values of x, let $f(x) = \cos\frac{1}{x}$. The function $\phi(x)$, associated with $f(x)$, agrees with $\cos\frac{1}{x}$ at every point except $x = 0$, where $\phi(x)$ is represented by $(-1, +1)$. The measure of discontinuity k_f is zero except at $\frac{1}{\pi}, \frac{1}{2\pi}, \frac{1}{3\pi}, \ldots$, where $k_f = 1$, and at $x = 0$, where $k_f = 2$; the measure k_ϕ vanishes everywhere except at $x = 0$, where $k_\phi = 2$. The function $\chi(x)$ vanishes except at $\frac{1}{\pi}, \frac{1}{2\pi}, \ldots$, where it is 1; it vanishes at $x = 0$, but is discontinuous at that point.

2.† Let $f(x)$ vanish except at the points $x = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \frac{1}{n}, \ldots$, where $f(x) = 1$. The function $\phi(x)$ is everywhere zero, and thus k_ϕ is everywhere zero. The function $\phi_1(x)$ is everywhere zero, and k_f is zero except at $0, \frac{1}{2}, \frac{1}{3}, \ldots \frac{1}{n}, \ldots$, where $k_f = 1$. In this case $f(x) - \phi_1(x) = f(x)$.

3.† A point-wise discontinuous function $f(x)$ can be constructed‡ such that the function $\phi(x)$ may have, at a point x_0, the values belonging to a prescribed closed set G, in accordance with Bettazzi's theorem (§ 240). If G be unenumerable, choose an enumerable set G_1, dense in G, and let $g_1, g_2, g_3, \ldots$ be the points of G_1. Take a set of intervals $\{\delta_n\}$, where δ_n is

$$\left(x_0 + \frac{1}{2^n}, \; x_0 + \frac{1}{2^{n-1}}\right),$$

and define $f(x)$ as follows:—in $\delta_1, \delta_3, \delta_5, \delta_7, \ldots$ let $f(x) = g_1$; in $\delta_2, \delta_6, \delta_{10}, \ldots$ let $f(x) = g_2$; in $\delta_4, \delta_{12}, \delta_{20}, \ldots$ let $f(x) = g_3$; in general $f(x) = g_n$, in the first free interval, and in every second of the following free intervals; further let $f(x) = g_1$, for $x \leq x_0$. The function $f(x)$

---

* This definition is not in complete agreement with that of Schoenflies, see *Bericht*, vol. I, p. 134, to whom the term is due. Some erroneous statements of Schoenflies, in this connection, were pointed out and corrected by Hahn; see *Monatshefte f. Math.* vol. XVI (1905), p. 312.

† See Hahn, *loc. cit.*

‡ This is contrary to a statement of Schoenflies, see *Bericht*, vol. I, p. 135.

is point-wise discontinuous, the points of discontinuity being x_0 and the points $x_0 - \frac{1}{2n}$. The function $\phi(x)$ has two values at the points $x_0 + \frac{1}{2n}$; and at x_0 it has all the values of G_1, and therefore all those of G.

FUNCTIONS OF BOUNDED VARIATION

243. Let us suppose that the interval (a, b), for which a function $f(x)$ is defined, is divided into a number of parts

$$(x_0, x_1)\ (x_1, x_2) \ldots (x_{r-1}, x_r) \ldots (x_{n-1}, x_n), \text{ where } x_0 = a,\ x_n = b;$$

these parts forming a net with closed meshes.

Consider the sum

$$|f(x_1) - f(x_0)| + |f(x_2) - f(x_1)| + \ldots + |f(x_n) - f(x_{n-1})|,$$

or

$$\sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})|.$$

If the function be such that the sum $\sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})|$ *is less than some fixed positive number, for all possible nets, the function* $f(x)$ *is said to be of bounded* variation (à variation bornée) in the interval* (a, b).

The function $f(x)$ being of bounded variation, there must exist a number which may be denoted by $V_a^b f(x)$, such that, for every net,

$$\sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})| \leq V_a^b f(x),$$

and such that, if ϵ be an arbitrarily chosen positive number, a net

$$(x_0, x_1), (x_1, x_2), \ldots (x_{n-1}, x_n),$$

where $x_0 = a$, $x_n = b$, can be so determined that

$$\sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})| > V_a^b f(x) - \epsilon.$$

The number $V_a^b f(x)$ which satisfies these conditions is defined to be the *total variation* of $f(x)$ in the interval (a, b).

In case the function is not of bounded variation in (a, b), $V_a^b f(x)$ may be regarded as infinite, and it will be possible to determine a net, for which

$$\sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})| > N,$$

where N is an arbitrarily chosen positive number.

A function of bounded variation has only ordinary points of discontinuity.

Accordingly the set of points of discontinuity of such a function is enumerable.

---

* See Jordan's *Cours d'Analyse*, vol. I, p. 55.

To see that this statement is correct, let us suppose that, at a point ξ, in (a, b), the upper and lower limits of the function on one side, say the right, have different values; thus let $\overline{f(\xi+0)}$ be $> \underline{f(\xi+0)}$. If η be an arbitrarily chosen positive number, in the neighbourhood of ξ there is an indefinitely great number of points at which $f(x) > \overline{f(\xi+0)} - \eta$, and also such a set at which $f(x) < \underline{f(\xi+0)} + \eta$. It follows that an indefinitely great number of non-overlapping intervals exists, in each of which the absolute value of the difference of the functional values at the end-points is

$$> \overline{f(\xi+0)} - \underline{f(\xi+0)} - 2\eta,$$

which is a positive number κ, provided η be chosen to be less than

$$\tfrac{1}{2}\{\overline{f(\xi+0)} - \underline{f(\xi+0)}\}.$$

A set of m intervals can be so chosen that, if their end-points belong to those of a net, their contribution to $\sum\limits_{r=1}^{r=n} | f(x_r) - f(x_{r-1}) |$ is $> m\kappa$. Since m may be taken as large as we please, it is clear that the function cannot have bounded variation.

It is obvious that $f(b) - f(a) = \sum\limits_{r=1}^{r=n} \{f(x_r) - f(x_{r-1})\}$; now let the terms in the sum on the right-hand side be divided into two parts, those for which $f(x_r) - f(x_{r-1})$ is positive, and those for which it is negative. Thus

$$f(b) - f(a) = \Sigma_1 \{f(x_r) - f(x_{r-1})\} + \Sigma_2 \{f(x_r) - f(x_{r-1})\}$$

where, in Σ_1, those values of r are taken for which $f(x_r) - f(x_{r-1})$ is positive, and in Σ_2, those terms for which it is negative.

If the function is of bounded variation, the value of

$$\Sigma_1 \{f(x_r) - f(x_{r-1})\} - \Sigma_2 \{f(x_r) - f(x_{r-1})\}$$

cannot exceed $V_a^b f(x)$. Thus

$$\Sigma_1 \{f(x_r) - f(x_{r-1})\} \leqq \tfrac{1}{2}\{V_a^b f(x) + f(b) - f(a)\},$$

and

$$-\Sigma_2 \{f(x_r) - f(x_{r-1})\} \leqq \tfrac{1}{2}\{V_a^b f(x) - f(b) + f(a)\}.$$

It follows that the two sums

$$\Sigma_1 \{f(x_r) - f(x_{r-1})\}, \quad \Sigma_2 \{f(x_r) - f(x_{r-1})\}$$

are both bounded, for all nets fitted on to (a, b).

Choosing a net fitted on to (a, b), for which

$$\sum_{r=1}^{r=n} | f(x_r) - f(x_{r-1}) | > V_a^b f(x) - \epsilon,$$

we see that

$$\Sigma_1 \{f(x_r) - f(x_{r-1})\} > \tfrac{1}{2}\{V_a^b f(x) + f(b) - f(a) - \epsilon\}$$

$$-\Sigma_2 \{f(x_r) - f(x_{r-1})\} > \tfrac{1}{2}\{V_a^b f(x) - f(b) + f(a) - \epsilon\}.$$

It follows that the numbers

$$\tfrac{1}{2}\{V_a^b f(x) + f(b) - f(a)\},\ \tfrac{1}{2}\{V_a^b f(x) - f(b) - f(a)\}$$

are the upper boundaries of

$$\Sigma_1 \{f(x_r) - f(x_{r-1})\},\ -\Sigma_2 \{f(x_r) - f(x_{r-1})\}$$

for all possible nets fitted on to the interval (a, b).

Denoting these upper boundaries by $P_a^b f(x)$, $N_a^b f(x)$ respectively, we have

$$V_a^b f(x) = P_a^b f(x) - N_a^b f(x),$$

$$f(b) - f(a) = P_a^b f(x) - N_a^b f(x).$$

The numbers $P_a^b f(x)$, $-N_a^b f(x)$ may be called the *total positive variation*, and the *total negative variation*, of $f(x)$ in (a, b).

The following definition* of a function of bounded variation is equivalent to the one given above:

If any set of non-overlapping intervals, finite, or indefinitely great, in number, be defined in (a, b), and the sum, or limiting sum, of $|f(\beta) - f(\alpha)|$, for all the intervals (α, β) of the set, be less than some fixed number, independent of the particular set of intervals, the function $f(x)$ is said to have bounded variation in (a, b).

In the first place, we see that a function which satisfies this definition must satisfy the definition employed above. For we may take the set of intervals to be those of any net fitted on to (a, b). Conversely, if the first definition is satisfied by $f(x)$, we may arrange any infinite set of non-overlapping intervals in order of descending magnitude of $|f(\beta) - f(\alpha)|$ and any finite part of the set containing the first m of the intervals is part of a net in (a, b); and thus $\Sigma\,|f(\beta) - f(\alpha)|$, for these m intervals, is less than $V_a^b f(x)$. As this holds whatever value m may have, we see that the second definition is satisfied by $f(x)$.

243^1. *If the function $f(x)$ be not of bounded variation (a, b), a system of nets can be fitted on to (a, b) such that the sum $S_D \equiv \sum_{r=1}^{r=n} |f(x_r) - f(x_{r-1})|$, for a net $(a, x_1, x_2, \ldots, b)$, increases indefinitely, as D consists successively of the nets $D_1, D_2, \ldots D_n, \ldots$, of the system of nets.*

If $\{A_n\}$ is a divergent sequence of increasing positive numbers, a net D_1 can be so determined that S_{D_1} is $> A_1$: and a net Δ_1 can be so determined that $S_{\Delta_1} > A_2$; taking D_2 to be the net (D_1, Δ_1) obtained by superimposing the two nets D_1, Δ_1, we have $S_{D_2} \geqq S_{\Delta_1} > A_2$. We proceed to form a net D_3 or (D_2, Δ_2), where Δ_2 is such that $S_{\Delta_2} > A_3$; then $S_{D_3} > A_3$; and so on. Thus the required system of nets $D_1, D_2, D_3, \ldots$ is determinate.

If $f(x)$ is not of bounded variation in (a, b), and if a net D be fitted on to (a, b), in one at least of the meshes of D the function $f(x)$ is not of bounded variation.

---

* See W. H. Young, *Quarterly Journal*, vol. XLII (1911), p. 57.

Consider the net D_n, for which $S_{D_n} > A_n$, employed in the last theorem. The net (D, D_n), obtained by superimposing D_n and D, is such that $S_{(D, D_n)} > A_n$. Now (D, D_n) consists of a sum of nets fitted on to the meshes of the net D; hence there must be at least one of the meshes of D, for which the total variation of $f(x)$ is $\geqq \frac{A_n}{m}$, where m is the number of meshes in D. This is applicable for each value of n; hence there is at least one mesh of D such that, for an infinite set of values of n, the total variation of $f(x)$ over that mesh is $\geqq \frac{A_n}{m}$. In such mesh of D the variation of $f(x)$ is unbounded.

Every absolutely continuous function is of bounded variation.

Let $\phi(x)$ be a function which is absolutely continuous in (a, b) (see § 218). If $\phi(x)$ is not of bounded variation in (a, b), a system of nets with closed meshes can be fitted on to (a, b) such that the sum

$$\sum_{r=1}^{r=m_n} \left| \phi(x_r^{(n)}) - \phi(x_{r-1}^{(n)}) \right|$$

for the net D_n increases indefinitely, as $n \sim \infty$. Since $\phi(x)$ is not of bounded variation in (a, b), there must be at least one mesh of D_1, say d_{p_1}, in which $\phi(x)$ is not of bounded variation; this follows from the preceding theorem. Further, d_{p_1} must contain at least one mesh d_{p_2}, of D_2, which has the same property; and so on. For a large enough value of n, d_{p_n} has a measure less than an arbitrarily chosen positive number. In d_{p_n}, for each value of n there is a set of intervals for which the sum of the absolute differences of $\phi(x)$ at the ends of an interval is $> A$, an arbitrarily chosen positive number. Hence a set of non-overlapping intervals, of total measure arbitrarily small, exists such that the sum of the absolute values of the differences of $\phi(x)$ at the ends of an interval is $> A$; and this is contrary to the hypothesis that $\phi(x)$ is absolutely continuous. Therefore $\phi(x)$ must be of bounded variation in (a, b).

If a function $f(x)$ is not of bounded variation in (a, b), there must be at least one point ζ, in (a, b), such that, in no neighbourhood of ζ, is $f(x)$ of bounded variation. For, consider a system of nets $\{D_n\}$; in D_1, there is at least one mesh d_{p_1} in which $f(x)$ has unbounded variation. In at least one mesh d_{p_2} of D_2, contained in d_{p_1}, the variation of $f(x)$ is unbounded. In this manner a sequence $d_{p_1}, d_{p_2}, \ldots$ of meshes, each of which is contained in the preceding one, is determined such that, in each of them, the variation is unbounded. This sequence defines a point ζ which has the required property; for any neighbourhood of ζ contains d_{p_n}, if n is sufficiently large. The point ζ may be said to be a *point of unbounded variation* of $f(x)$.

It is easily seen that:

For any function $f(x)$, of unbounded variation in (a, b), there exists a set of points of unbounded variation which is closed, and may be finite.

For a limiting point of the set of points of unbounded variation is such that any neighbourhood of it contains points of the set, and therefore contains intervals in which the function is of unbounded variation.

FUNCTION OF BOUNDED VARIATION EXPRESSED AS THE DIFFERENCE OF TWO MONOTONE FUNCTIONS

244. It is clear that, if x be any point in (a, b), a function $f(x)$ which has bounded variation in (a, b) has also bounded variation in (a, x): and consequently the positive, and the negative, variations of the function in (a, x) are also bounded.

The three numbers which represent these total variations may be denoted by $V_a^x f(x)$, $P_a^x f(x)$, $-N_a^x f(x)$; and they may be regarded as functions of x.

They satisfy the relations

$$V_a^x f(x) = P_a^x f(x) + N_a^x f(x),$$
$$f(x) - f(a) = P_a^x f(x) - N_a^x f(x).$$

It is clear that, if $x' > x$, we have

$$P_a^{x'} f(x) \geq P_a^x f(x), \quad N_a^{x'} f(x) \geq N_a^x f(x);$$

and thus the two functions $P_a^x f(x)$, $N_a^x f(x)$ are monotone non-diminishing functions, in (a, b). For simplicity, we denote these functions by $P(x)$, $N(x)$ respectively.

Since $\qquad f(x) = \{P(x) + f(a) + k\} - \{N(x) + k\}$

and $\qquad f(x) = \{k' - N(x)\} - \{k' - P(x) - f(a)\}$,

where k, k' are arbitrarily chosen numbers, we see that:

Every function of bounded variation can be expressed, in an indefinitely great number of ways, as the difference of two bounded monotone functions, both of which are non-diminishing, or both of which are non-increasing.

This theorem expresses the cardinal property of functions of bounded variation. It is of great value in Analysis, as, by means of it, properties of monotone functions can be immediately extended to the wide class of functions of bounded variation.

The converse theorem holds that:

The difference of two bounded monotone functions, both non-diminishing, or both non-increasing, is a function of bounded variation.

For, consider such a function $f(x) = P(x) - N(x)$, where $P(x)$, $N(x)$ are non-diminishing; then, in any interval (α, β), we have

$$|f(\beta) - f(\alpha)| \leq \{P(\beta) - P(\alpha)\} + \{N(\beta) - N(\alpha)\}.$$

Hence, in any net fitted on to (a, b), the sum of the absolute values of the differences of the values of $f(x)$ at the ends of the meshes (α, β) is

$$\leq \Sigma \{P(\beta) - P(\alpha)\} + \Sigma \{N(\beta) - N(\alpha)\}$$
$$\leq P(b) - P(a) + N(b) - N(a);$$

and thus the theorem is established.

It has now been shewn that it is necessary and sufficient, in order that a function may be of bounded variation in the interval for which it is defined, that it can be expressed as the difference of two bounded monotone non-diminishing (or non-increasing) functions.

The sum, difference, or product, of two functions of bounded variation is also of bounded variation.

If $f_1(x) = P_1(x) - N_1(x), f_2(x) = P_2(x) - N_2(x)$, the theorem follows at once from the relations

$$f_1(x) + f_2(x) = \{P_1(x) + P_2(x)\} - \{N_1(x) + N_2(x)\},$$
$$f_1(x) - f_2(x) = \{P_1(x) + N_2(x)\} - \{P_2(x) + N_1(x)\},$$
$$f_1(x) f_2(x) = \{P_1(x) P_2(x) + N_1(x) N_2(x)\} - \{P_1(x) N_2(x) + P_2(x) N_1(x)\}.$$

244^1. *If the interval (a, b) be divided into two or more parts, a function which is of bounded variation in each part is of bounded variation in (a, b). The total variation in (a, b) is the sum of the total variations in the parts.*

It is clearly sufficient to consider the case in which (a, b) is divided into two parts (a, c), (c, b).

If $f(x)$ is of bounded variation in (a, c) it is expressible in that interval as $\phi_1(x) - \psi_1(x)$, where $\phi_1(x)$, $\psi_1(x)$ are the monotone non-diminishing functions $f(a) + P_a{}^x f(x)$, $N_a{}^x f(x)$. In the interval (c, b) it can be expressed similarly in the form $\phi_2(x) - \psi_2(x)$, where $\phi_2(x)$, $\psi_2(x)$ denote

$$\lambda + f(c) + P_c{}^x f(x), \quad \lambda + N_c{}^x f(x),$$

respectively. We have $f(c) = \phi_1(c) - \psi_1(c) = \phi_2(c) - \psi_2(c)$; and we may choose λ so that $\psi_1(c) = \psi_2(c)$. It then follows that $\phi_1(c) = \phi_2(c)$; and thus $\phi_1(x)$ in (a, c), $\phi_2(x)$ in (c, b) define a single monotone function $\phi(x)$ in (a, b); similarly the monotone function $\psi(x)$ is defined to be equal to $\psi_1(x)$ in (a, c) and $\psi_2(x)$ in (c, b). Since $f(x)$ is, in (a, b), the difference of two monotone non-diminishing functions, it is of bounded variation in (a, b).

The total variation in (a, c) is $\{\phi_1(c) - \phi_1(a)\} + \{\psi_1(c) - \psi_1(a)\}$; that in (c, b) is $\{\phi_2(b) - \phi_2(c)\} + \{\psi_2(b) - \psi_2(c)\}$, and the sum of the two is $\{\phi_2(b) - \phi_1(a)\} + \{\psi_2(b) - \psi_1(a)\}$, which is the total variation in (a, b).

FUNCTIONS OF BOUNDED TOTAL FLUCTUATION

245. If, in the net $(x_0, x_1), (x_1, x_2), \dots (x_{n-1}, x_n)$, where $x_0 = a$, $x_n = b$, we take the sum of the fluctuations in the closed meshes of the net, viz. $\sum\limits_{r=1}^{r=n} (U_r - L_r)$, where U_r, L_r are the upper and lower boundaries of $f(x)$ in the closed mesh (x_{r-1}, x_r), then if this sum does not exceed some fixed positive number, independent of the particular net, the function $f(x)$ is said to have *bounded total fluctuation* in the interval (a, b). If $F_a^b f(x)$ be the number such that, for every net,

$$\sum_{r=1}^{r=n} (U_r - L_r) \leq F_a^b f(x),$$

and such that a net can be fitted on to (a, b) for which

$$\sum_{r=1}^{r=n} (U_r - L_r) > F_a^b f(x) - \epsilon,$$

where ϵ is an arbitrarily chosen positive number, then $F_a^b f(x)$ is termed[*] the *total fluctuation* of $f(x)$ in (a, b); and it is finite for a function of bounded total fluctuation. Every discontinuity of $f(x)$ is then ordinary (see § 243).

It will appear that the class of functions of bounded total fluctuation is identical with that of functions of bounded variation.

In order that a function may have bounded total fluctuation it is sufficient that $\Sigma\,(U_r - L_r)$*, when taken over the successive nets of a given system of nets, should have a definite limit. Moreover the limit for the system of nets is*

$$\leq F_a^b f(x), \text{ and } \geq \tfrac{1}{2} F_a^b f(x).$$

This limit may be called the total fluctuation of $f(x)$ for the particular system of nets.

It is clear that, if an interval (α, β) be divided into two parts (α, γ), (γ, β), the sum of the fluctuations in (α, γ), (γ, β) cannot be less than the fluctuation in (α, β). Hence, if we consider the successive nets $D_1, D_2, \dots$ of a system of nets, the sum of the fluctuations in the meshes of D_m is $\leq$ the sum in the meshes of D_{m+1}. Thus, as we proceed through the successive nets of the system, the sums of the fluctuations in the meshes of a net form a sequence of non-diminishing numbers which have a finite upper limit, unless they increase indefinitely. Assume the former to be the case. Let us now consider a second system of nets $D_1', D_2', \dots D_m', \dots$. Suppose η to be a number less than the breadth of all the meshes of D_m'; and let the two nets D_r, D_m' be superimposed to form a net (D_r, D_m'). We may suppose r so large that D_r contains no mesh of breadth $> \eta$. A mesh d_s, of D_r, with fluctuation F_r, will contain not more than two meshes of the net (D_r, D_m'), with fluctuations F_{r_1}, F_{r_2}; we have then

---

[*] See Study, *Math. Annalen*, vol. XLVII (1896), p. 298. It was shewn by Study that the definitions of functions of bounded total fluctuation, and of bounded variation, are equivalent to one another.

$$F_r \leqq F_{r_1} + F_{r_2} \leqq 2F_r.$$

Now every mesh of D_m' is made up of meshes of D_r and of meshes of (D_r, D_m'). Therefore the sum of the fluctuations in the meshes of D_m' is $\leqq$ the sum of the fluctuations in the undivided meshes of D_r together with the sum of the fluctuations such as F_{r_1}, F_{r_2}. It follows that the sum of the fluctuations in the meshes of D_m' cannot exceed twice the sum of the fluctuations in the meshes of D_r. Since a number r can be determined so as to correspond to each number m, we see that the limit of the sum of the fluctuations in D_m', as $m \sim \infty$, cannot exceed twice the limit of the sum of the fluctuations in D_r, as $r \sim \infty$. If X' be the former limit, and X the latter, we have $X' \leqq 2X$. Thus, if X is finite, for any system of nets, it is finite for any other system of nets.

Since a net can be so chosen that the sum of the fluctuations in its meshes is less than $F_a^b f(x)$ by less than an arbitrarily chosen number, a system of nets can be so chosen that the limit of the sums for the nets is $F_a^b f(x)$. It thus follows that, for any other system of nets, the limit of the sum of the fluctuations is $\geqq \frac{1}{2} F_a^b f(x)$, and $\leqq F_a^b f(x)$.

246. Since the absolute value of the difference of the functional values at the end-points of an interval (α, β) cannot exceed the fluctuation in the closed interval, it is clear that the sum $\sum_{r=1}^{r=n} | f(x_r) - f(x_{r-1}) |$ cannot exceed the sum of the fluctuations in the intervals. It follows that, if a function is of bounded total fluctuation, it is also of bounded variation.

If a function have at no point an external saltus, the total fluctuation, and the total variation, have one and the same value, when estimated as the limiting values over a system of nets; and that value is independent of the particular system of nets.

In accordance with this theorem, $V_a^b f(x)$ and $F_a^b f(x)$ are equal to one another, and are the limits of the sums

$$\Sigma \, | f(x_r) - f(x_{r-1}) |, \quad \Sigma \, (U_r - L_r),$$

over a net D_n, of a system of nets, as $n \sim \infty$.

To prove the theorem, let $V^{(r)}, F^{(r)}$ denote $| f(x_r) - f(x_{r-1}) |$, $(U_r - L_r)$ for a mesh d_r, of a net D_r. We have clearly $V^{(r)} \leqq F^{(r)}$. It will be shewn that for the net D_{r+s}, where s is sufficiently large, $\Sigma V^{(r+s)} \geqq \Sigma F^{(r)}$, the summation referring to all the meshes of D_{r+s} that are in the mesh d_r, of the net D_r. In the case in which U_r, L_r are the functional values at the ends of d_r, we have $U_r - L_r = | f(x_r) - f(x_{r-1}) |$. When this is not the case, one at least of $f(x_r), f(x_{r-1})$ is between U_r and L_r. If ϵ be an arbitrarily chosen positive number, there are two points of d_r, say ξ_1, ξ_2, for one of which the functional value is $> U_r - \epsilon$, and for the other $< L_r - \epsilon$. There must exist a neighbourhood of ξ_1, on one side at least, at every point of which the functional value is $> U_r - 2\epsilon$: for otherwise the functional limits at ξ_1 would be $< U_r - \epsilon$. on both sides, and this cannot be the case if there is no external saltus at ξ_1. It follows that, if s be sufficiently large, there must be an end-point of a mesh of D_{r+s} at which the functional value is $> U_r - 2\epsilon$. Similarly we see that, if s be sufficiently large, there is an end-point of a mesh of D_{r+s} at which the functional value is $< L_r + 2\epsilon$; and one at least of these end-points is not an end-point of the mesh d_r. The number s can accordingly be so chosen that, for a certain number of meshes of the net D_{r+s} which are in the mesh d_r of D_r, the sum of the absolute values of the differences of the functional values is $> U_r - L_r - 4\epsilon$. Thus the number s can be so chosen that the sum of the absolute values of the differences of the functions for all those meshes of D_{r+s} that are in the mesh d_r is $> U_r - L_r - 4\epsilon - W$, where W is the absolute difference between the functional value at one of the end-points of d_r and the functional value at one of those end-points of a mesh of D_{r+s} at which it is $> U_r - 2\epsilon$ or $< L_r - 2\epsilon$. If ϵ be so chosen that $W > 4\epsilon$, (which can be done, because the functional value at the end-point of d_r in question is not equal to U_r or L_r), we see that the sum of the absolute differences of the functional values in those meshes of D_{r+s} that are in d_r is $> U_r - L_r$. Moreover s can be so chosen that this holds good for each mesh d_r, of the net D_r; and thus, for such value of s, we have $V^{(r+s)} > F^{(r)}$. This inequality, combined with the inequality $V^{(r)} \leq F^{(r)}$, shews that

$$\lim_{r\sim\infty} V^{(r)} = \lim_{r\sim\infty} V^{(r+s)} \geq \lim_{r\sim\infty} F^{(r)},\quad \lim_{r\sim\infty} V^{(r)} \leq \lim_{r\sim\infty} F^{(r)},$$

and thus that the two limits $\lim\limits_{r\sim\infty} V^{(r)}$, $\lim\limits_{r\sim\infty} F^{(r)}$ have the same value, finite or infinite.

In order to complete the proof of the theorem it must be shewn that the limit of the sum of the fluctuations in the meshes of a net belonging to a system of nets is the same for all systems of nets; there being as before no external saltus at any point.

If possible, let the limit for a particular system of nets have a value G, less than $F_a^b f(x)$. A net can be defined, for which the sum of the fluctuations in the m meshes is $> G$; say $G + \alpha$. Let this net D' be superimposed on the net D_r, of the system of nets; we may suppose r so great that not more than one of the end-points of the meshes of D' is in any one of the meshes of D_r. Let us suppose that one of these end-points is x, and that it falls in the mesh δ, of D_r, dividing it into two parts δ' and δ''; let $F(\delta)$ be the fluctuation in δ. Since there is no external saltus, if δ, and consequently δ' and δ'', are sufficiently small, we have

$$F(\delta) = |f(x+0) - f(x-0)| + G_1,$$

where G_1 is less than the arbitrarily chosen number η. Moreover, δ' and δ'' being sufficiently small,

$$F(\delta') = |f(x) - f(x-0)| + G_2, \text{ where } G_2 < \eta,$$
$$F(\delta'') = |f(x) - f(x+0)| + G_3, \text{ where } G_3 < \eta.$$

Now r may be chosen so great, and consequently all the meshes of D_r so small, that these conditions are satisfied for all the meshes of D'. We have then $F(\delta') + F(\delta'') - F(\delta) = G_2 + G_3 - G_1$, since $f(x)$ is between $f(x+0)$ and $f(x-0)$. Hence the sum

$$\Sigma F(\delta') + \Sigma F(\delta'') - \Sigma F(\delta) < 2(m-1)\eta,$$

the summation being taken for all those meshes which contain one of the $m-1$ points x; moreover $\Sigma F(\delta') + \Sigma F(\delta'') - \Sigma F(\delta) \geqq 0$. Therefore the sum of the fluctuations in the net obtained by superimposing D' and D_r is $< G + 2(m-1)\eta$. But the sum of these fluctuations is certainly $\geqq G + \alpha$, and since η is arbitrarily small and independent of m, these two conditions are incompatible with one another; thus α must be zero. It follows that G cannot be less than $F_a^b f(x)$. Thus, in every system of nets, the limit of the sum of the fluctuations in the meshes of a net is $F_a^b f(x)$. It then follows, from the first part of the theorem, that the limit of the sum of the absolute differences of the functional values for the meshes of any net is $V_a^b f(x)$, or $F_a^b f(x)$.

In case there be points at which there is an external saltus, the preceding proof can still be applied to shew that the limit of the sums of the fluctuations in two systems of nets has the same value for the two systems, provided no point at which there is an external saltus be an end-point of a mesh in either system. Moreover, in case there be an external saltus at the point x, the value of $F(\delta)$ is either $|f(x) - f(x+0)| + G_1$ or $|f(x) - f(x-0)| + G_1$, and we see that

$$F(\delta') + F(\delta'') - F(\delta) = G_2 + G_3 - G_1 + s(x),$$

where $s(x)$ denotes the external saltus at x, and is equal to the smaller of the two numbers $|f(x) - f(x+0)|$, $|f(x) - f(x-0)|$.

It is clear that, for a function with bounded total fluctuation, the sums of the saltuses on the right, and on the left,

$$\Sigma\,|f(x+0) - f(x)|, \quad \Sigma\,|f(x-0) - f(x)|,$$

for all the points of discontinuity, are finite, and there can therefore be only a finite number of points at which the external saltus exceeds an arbitrarily chosen number β. If we take these points, say $n-1$ in number, to be the points x above, of D', we see that the sum of the fluctuations in the meshes of the net obtained by superimposing D' and D_r is

$$\Sigma F(\delta) + S_\beta + \gamma,$$

where S_β is the sum of those external saltuses, all of which are greater than β; and γ is the sum of the $n-1$ numbers $G_2 + G_3 - G_1$, and is therefore arbitrarily small. It thus appears that the total fluctuation for a system of nets such that no end-point of any of the nets is a point at which there is an external saltus is the number μ ($\geq \frac{1}{2} F_a^b f(x)$), the lower boundary of the limiting sum for all systems of nets. If those points at which the external saltus is $> \beta$ be end-points of meshes of nets of the system, the limit of the sum of the fluctuations in a net of the system is $\mu + S_\beta$. If a sequence of descending values be given to β, the sum S_β converges to a fixed number, which is the sum, or the limiting sum, of all the external saltuses. Thus we have the following theorem

If a function with bounded total fluctuation be such that there are points at which the function has an external saltus, then the difference $F_a^b f(x) - \mu$ *between the upper and lower boundaries of the limiting sum of the fluctuations in the nets of a system is equal to the sum of the external saltuses. If a system of nets be such that no point at which there is an external saltus is an end-point of any mesh of any of the nets, for such a system the limiting sum of the fluctuations in the meshes of a net is* μ. *If however every point at which there is an external saltus is an end-point of a mesh of nets of the system, the limiting sum is then* $F_a^b f(x)$.

247. It has now been shewn that, for a function without points at which there is an external saltus, the total variation and the total fluctuation over any system of nets are identical, being both finite, or both infinite, and that they are independent of the particular system of nets employed.

If $f(x)$ have an external saltus at each point of some set S, the total variation of $f(x)$, when extended over a system of nets such that no point of S is an end-point of any mesh, is identical with the total fluctuation of that function $\phi(x)$ which differs from $f(x)$ only in having the functional values at the points of S so altered that the external saltus is at every point of S removed. The total fluctuation of $\phi(x)$ is obtained by removing from μ the total fluctuation of $f(x)$ over such a system of nets, that is the number $F_a^b f(x) - \mu$ which is the sum of the external saltuses. Thus the total fluctuation of $\phi(x)$ is $2\mu - F_a^b f(x)$. If, on the other hand, we employ a system of nets such that every point of S is an end-point of meshes of the nets, the total variation and the total fluctuation of $f(x)$ have the common value $F_a^b f(x)$.

It thus appears that, *for a function of bounded total fluctuation, which has points with an external saltus, the total variation over a system of nets is* $F_a^b f(x)$, *or* $2\mu - F_a^b f(x)$, *or has some value between these two numbers, according to the particular system of nets employed.*

The necessary and sufficient conditions that a function $f(x)$ may have bounded variation and bounded fluctuation in (a, b) are that:

(1). The points of discontinuity must all be of the first species, *i.e.* $f(x+0)$, $f(x-0)$ must everywhere exist.

(2). The sum of the absolute values of the external saltuses must be finite.

(3). A system of nets must exist for which the sum of the absolute values of the differences of the functional values at the end-points of the meshes of a net must be less than some fixed number, for all the nets of the system.

248. The following theorem will now be established:

The necessary and sufficient condition that a function $f(x)$ is of bounded total fluctuation is that it can be expressed as the difference of two bounded monotone functions, which are either both non-increasing or both non-diminishing.

A comparison of this theorem with that in § 244, for functions of bounded variation, gives a second proof that the two classes of functions are identical.

To prove the theorem; let F_a^x be the upper boundary of the total fluctuation in the interval (a, x). We then see that

$$f(x+h) - f(x) \leqq F_x^{x+h} \leqq F_a^{x+h} - F_a^x;$$

it follows that $F_a^x - f(x)$ is a monotone non-diminishing function The function $F_a^x + f(x)$ has the same property, since

$$f(x+h) - f(x) \geqq -F_x^{x+h} \geqq F_a^x - F_a^{x+h}.$$

Therefore, if $\phi_1(x) = \frac{1}{2}\{F_a^x + f(x)\}$, $\phi_2(x) = \frac{1}{2}\{F_a^x - f(x)\}$, the function $f(x)$ can be expressed as the difference of the two non-diminishing monotone functions $\phi_1(x)$, $\phi_2(x)$.

Conversely, let $f(x) = \phi_1(x) - \phi_2(x)$; where $\phi_1(x)$, $\phi_2(x)$ are bounded non-diminishing monotone functions. In any interval, the fluctuation of $f(x)$ cannot exceed the sum of the fluctuations of $\phi_1(x)$ and $\phi_2(x)$. It then follows that the total fluctuation of $f(x)$ in (a, b) cannot exceed

$$[\phi_1(b) - \phi_1(a)] + [\phi_2(b) - \phi_2(a)].$$

EXAMPLES

1.* The function defined by $f(x) = x \sin\frac{1}{x}$, $f(0) = 0$, is not of bounded total fluctuation in the interval $(0, 1/\pi)$, although it is continuous in the interval. For, in the interval $\left(\frac{1}{\overline{r+1}\,\pi}, \frac{1}{r\pi}\right)$, $\sin\frac{1}{x}$ attains the value $(-1)^r$, and thus the fluctuation in this interval is at least equal to $1/(r+\frac{1}{2})\pi$. The total fluctuation in the interval $\left(\frac{1}{s\pi}, \frac{1}{\pi}\right)$ is at least $\frac{1}{\pi}\left\{\frac{1}{1+\frac{1}{2}} + \frac{1}{2+\frac{1}{2}} + \dots + \frac{1}{s-1+\frac{1}{2}}\right\}$ or $\frac{2}{\pi}\left(\frac{1}{3} + \frac{1}{5} + \dots + \frac{1}{2s-1}\right)$, and it is well known that this increases without limit when s is indefinitely increased; therefore the total fluctuation in $(0, 1/\pi)$ is not finite.

* Lebesgue, *Leçons sur l'intégration* (1904), p. 56.

2 [*] The function defined by $f(x) = x^2 \sin \frac{1}{x^2}$, $f(0) = 0$, is continuous in any interval containing $x = 0$, and is everywhere differentiable, but is not of bounded total fluctuation.

3.[†] The function defined by $f(x) = x^2 \sin(x^{-\frac{4}{3}})$, $f(0) = 0$, is of bounded total fluctuation in the interval $(0, 1/\pi^{\frac{3}{4}})$. In the interval $\left(\frac{1}{\{(r+1)\pi\}^{\frac{3}{4}}}, \frac{1}{(r\pi)^{\frac{3}{4}}}\right)$, the function has a single maximum, or else a single minimum, and the absolute value of the function at this point is at most $1/(r\pi)^{\frac{3}{2}}$. The total fluctuation in $(0, 1/\pi^{\frac{3}{4}})$ cannot exceed $2\sum_1^\infty \frac{1}{(r\pi)^{\frac{3}{2}}}$, which is finite.

RESOLUTION OF A FUNCTION OF BOUNDED VARIATION

249. If $f(x)$ have bounded variation in the interval (a, b), it can be expressed by $f(a) + P(x) - N(x)$, where $P(x)$, $-N(x)$ are the positive, and the negative, variation respectively of $f(x)$ in the interval (a, x). Let $\xi_1, \xi_2, \ldots \xi_n, \ldots$ be the points of discontinuity of $P(x)$, and let $s(x)$ denote the sum

$$\sum_x \{P(\xi + 0) - P(\xi - 0)\},$$

where the summation is taken for those points ξ that are in the closed interval (a, x). Thus $s(x)$ is a bounded non-diminishing monotone function. Since the sum $\sum_b \{P(\xi + 0) - P(\xi - 0)\}$ is convergent, the sum can be divided into two parts, the one a finite sum

$$\sum_{b,m} \{P(\xi + 0) - P(\xi - 0)\}, \text{ and the other } \sum_{b,m+1}^{\infty} \{P(\xi + 0) - P(\xi - 0)\},$$

where m is chosen so great as to include the indices of all those points of the finite set at which $P(\xi + 0) - P(\xi - 0) \geqq \epsilon$ and is also so great that the second sum is $< \epsilon$; where ϵ is an arbitrarily chosen positive number.

We also have $s(x) = \sum_{x,m} \{P(\xi + 0) - P(\xi - 0)\} - \eta(x)$; where, in the first sum, only those of the set $\xi_1, \xi_2, \ldots \xi_m$ occur which are in the closed interval (a, x) and where $\eta(x) \leqq \eta(b) < \epsilon$. Let us consider the function $P(x) - s(x)$, or $P(x) - \sum_{x,m} \{P(\xi + 0) - P(\xi - 0)\} - \eta(x)$. At a point x, at which $P(x)$ has a saltus $< \epsilon$, the saltus of $P(x) - s(x)$ is $< 2\epsilon$; for the saltus of $\eta(x)$ cannot exceed ϵ, and the function $\sum_{x,m} \{P(\xi - 0) - P(\xi - 0)\}$ is continuous. At a point of the finite set at which $P(x)$ has a saltus $\geqq \epsilon$, $P(x) - \sum_{x,m} [P(\xi + 0) - P(\xi - 0)]$ is continuous. Therefore the saltus of $P(x) - s(x)$ is everywhere $< 2\epsilon$. Since ϵ is arbitrary, it follows that $P(x) - s(x)$ is continuous in the closed interval (a, b). A similar result holds for the function $N(x)$. Therefore $f(x)$ is the sum of a continuous function $\phi(x)$, of bounded variation in (a, b), and of a function $s(x) - s'(x)$, which may be denoted by $\chi(x)$. The function $\chi(x)$ is of bounded variation, and has the same points of discontinuity as $f(x)$. We have

---

[*] Lebesgue, *Annali di Mat.* (3 A), vol. VII (1902), p. 270.

[†] Lebesgue, *Leçons sur l'intégration* (1904), p. 56.

$$f(x) = \phi(x) + \chi(x),$$

where $\chi(x)$ denotes

$$\sum_x [P(\xi + 0) - P(\xi - 0)] - \sum_x [N(\xi + 0) - N(\xi - 0)].$$

The result obtained may be stated as follows:

A function of bounded variation in the interval (a, b) is the sum of a continuous function of bounded variation and of a function of bounded variation with the same discontinuities as the given function, of which the total variation is the sum of the saltuses of the given function at its points of discontinuity.

RECTIFIABLE CURVES

250. Let t be a variable defined for all values in a continuous interval (t_0, t_1), and let $f_1(t)$, $f_2(t)$ be two single-valued and bounded functions of t, defined in the interval (t_0, t_1); the equations $x = f_1(t)$, $y = f_2(t)$ may be said to define an arc of a plane curve, in a generalized sense of the term. Let a net $\tau_0, \tau_1, \tau_2, \ldots \tau_n$, where $\tau_0 = t_0$, $\tau_n = t_1$, be fitted on to the interval (t_0, t_1), then the points on the curve corresponding to the end-points of the meshes of the net may be denoted by $P_0, P_1, \ldots P_n$; and we consider the unclosed polygon of which the sides are $P_0P_1, P_1P_2, \ldots P_{n-1}P_n$, inscribed in the arc of the curve. The length of the polygonal line $P_0P_1P_2 \ldots P_n$, measured by

$$\sum_{r=1}^{r=n} \{(x_r - x_{r-1})^2 + (y_r - y_{r-1})^2\}^{\frac{1}{2}},$$

depends upon the particular net that has been fitted on to (t_0, t_1), and may be described as the length of a polygonal arc inscribed in the arc of the curve.

If the arc be such that the lengths of all the inscribed polygonal arcs have a finite upper boundary, the curve is said to be rectifiable, and the length of the arc is defined to be the value of that upper boundary. Otherwise the length of the arc is regarded as infinite.

This general definition is due to Peano*. An earlier definition given by Jordan†, and in the case of an arc defined by an equation $y = f(x)$, by Scheeffer‡, is of a less general character, as it requires that the lengths of any sequence of polygonal arcs should converge to a fixed limit, independent of the particular sequence, provided the polygonal arcs correspond to a system of nets fitted on to (t_0, t_1). It will be shewn that when the arc is continuous, the definition of Peano is equivalent to that of Jordan and Scheeffer.

---

* *Rend. Lincei* (4), vol. VI, 1 (1890), p. 54. See also Lebesgue, *Ann. di Mat.* (3 A), vol. VII (1902), p. 282, where a more general definition is introduced.

† *Comptes Rendus*, vol. XCII (1881), p. 228; and *Cours d'Analyse*, vol. III, p. 594.

‡ *Acta Math.* vol. V (1884), p. 49. See also Study, *Math. Annalen*, vol. XLVII (1896), p. 298.

The following theorem will be established:

The necessary and sufficient condition that the arc defined by $x = f_1(t)$, $y = f_2(t)$, for $t_0 \leq t \leq t_1$, may be rectifiable is that the functions $f_1(t)$, $f_2(t)$ should be of bounded variation in the interval (t_0, t_1).

We see that

$$\frac{1}{\sqrt{2}}\{|x_r - x_{r-1}| + |y_r - y_{r-1}|\} \leq [(x_r - x_{r-1})^2 + (y_r - y_{r-1})^2]^{\frac{1}{2}}$$

$$\leq \{|x_r - x_{r-1}| + |y_r - y_{r-1}|\};$$

by taking the sums of all the expressions for $r = 1, 2, 3, \ldots n$, it is clear that the necessary and sufficient condition that the perimeter of the unclosed polygon should be less than some fixed positive number is that this should also hold as regards $\sum\limits_{r=1}^{r=n} |x_r - x_{r-1}|$, and $\sum\limits_{r=1}^{r=n} |y_r - y_{r-1}|$, and this is equivalent to the condition stated in the theorem.

From the known character of functions of bounded variation, it is seen that a curve that is rectifiable in the interval (t_0, t_1) is also rectifiable in any interval contained in (t_0, t_1).

251. In case the functions $f_1(t)$, $f_2(t)$ are both continuous in the interval (t_0, t_1), it will be shewn that the lengths of the polygonal arcs of any sequence, such that the greatest side of a polygon converges to zero, converge to the length l of the curve, as above defined, whenever the curve is rectifiable. To establish this result the following theorem will be proved.

If the continuous arc defined by $x = f_1(t)$, $y = f_2(t)$, $t_0 \leq t \leq t_1$, be rectifiable, and of length l, then, corresponding to an arbitrarily chosen positive number ϵ, a positive number d can be determined, so that the perimeter of any unclosed polygon inscribed in the arc is $> l - \epsilon$, provided all the sides of the polygon are $< d$.

The assumption that the arc is rectifiable, and of length l, implies that a net D_0 can be fitted on to the interval (t_0, t_1), such that the length of the corresponding polygonal arc inscribed in the curve is $> l - \frac{1}{2}\epsilon$. Let D be any other net fitted on to the interval (t_0, t_1). The net (D_0, D) obtained by superimposing the two nets D_0, D, is such that the corresponding polygonal arc has a length $> l - \frac{1}{2}\epsilon$. Since the functions $f_1(t)$, $f_2(t)$ are continuous, if η be a prescribed positive number, a number δ' can be determined so that, in any interval of t less than δ', the fluctuations of $f_1(t)$, $f_2(t)$ are both less than η. We shall suppose D so chosen that all its meshes are of breadth less than δ'. Let m be the number of meshes in D_0; we consider the excess of the length of the polygonal arc corresponding to (D_0, D) over that which corresponds to D. Consider an end-point t_u of the meshes of D_0, that is contained in the interior of a mesh of D; we can suppose the meshes of D so small that only one such point can be contained in a mesh of D, and we also suppose all the meshes of D to be less than δ'. If t_r, t_{r+1} be that mesh of D, we see that the excess of the sum of the chords joining t_r, t_u and (t_u, t_{r-1}) over that joining t_r, t_{r+1} is

$$< |x_r - x_u| + |x_{r+1} - x_u| - \frac{1}{\sqrt{2}}|x_r - x_{r+1}| + |y_r - y_u| \\ + |y_{r+1} - y_u| - \frac{1}{\sqrt{2}}|y_r - y_{r+1}|$$

and this is less than 4η.

We now see that the perimeter of the polygonal line which corresponds to D is $> l - \frac{1}{2}\epsilon - 4m\eta$.

If η is so chosen that $4m\eta < \frac{1}{2}\epsilon$, the length of the polygonal arc that corresponds to D is $> l - \epsilon$. In order that this may be the case, D has only to satisfy the condition that the maximum breadth of its meshes is less than some fixed number δ ($\leq \delta'$), and this condition will be satisfied if the greatest side of the corresponding polygonal line is less than some fixed number d.

If s denote the length of the arc corresponding to (t_0, t), where $t_0 < t \leqq t_1$, s is a continuous monotone function of t; and it is clear that x and y may be regarded as functions of s, say $x = \phi(s)$, $y = \psi(s)$. Since $\Delta x/\Delta s$, $\Delta y/\Delta s$ cannot be numerically greater than 1, all the derivatives of $\phi(s)$, $\psi(s)$ are bounded, and in the interval $(-1, 1)$. It will be shewn in § 298 that, for all values of s in the linear interval $(0, l)$, with the possible exception of those belonging to a set of measure zero, the differential coefficients $\frac{dx}{ds}, \frac{dy}{ds}$ exist, and are finite. The above theory is applicable, without essential change, to the case of a "curve" in three-dimensional space, defined by $x = f_1(t)$, $y = f_2(t)$, $z = f_3(t)$.

The case of an arc of a curve $y = f(x)$, defined for the interval (a, b), of x, may be considered as the particular case of the above which arises when t is identical with x. We see then that, *the necessary and sufficient condition that the arc defined by $y = f(x)$, $a \leqq x \leqq b$, may be rectifiable is that the function $f(x)$ is of bounded variation in the interval (a, b).*

EXAMPLES

1. In* the interval $(0, 1)$, let $f(x) = \frac{1}{2^{2n}}$, when x is any of the points $\frac{2r+1}{2^n}$, and let $f(x) = 0$, at all other points. In accordance with Peano's definition, the length of the arc is $1 + 2\sum_{n=1}^{\infty} \frac{2^{n-1}}{2^{2n}} = 2$. According to the definition of Jordan and Scheeffer, the arc is not rectifi-

---

* See Schoenflies, *Bericht*, vol. II, p. 243.

able; for a net can be fitted on to the interval (0, 1). such that all its end-points are represented by irrational values of x, except the points 0, 1; the length of the corresponding polygonal arc is 1.

2. An arc of the curve $y = x \sin \frac{1}{x}$ which contains the origin is not rectifiable: but a similar arc of the curve $y = x^2 \sin (x^{-4})$ is rectifiable.

THE VARIATION OF A FUNCTION OF BOUNDED VARIATION OVER A LINEAR SET OF POINTS

252. Let $\phi(x)$ be a bounded monotone non-diminishing function. defined for the interval (a, b), of x. If $\xi = \phi(x)$, the segment (a, b) has a functional image on a segment of which the end-points are $\phi(a)$. $\phi(b)$ respectively, on which the points ξ are represented. If $\phi(x)$ be continuous. the functional image consists of all the points of the interval $(\phi(a), \phi(b))$, of ξ; but if $\phi(x)$ be discontinuous at a point x', there are no points of the functional image in the closed interval $(\phi(x'-0), \phi(x'-0))$ except the single point $\phi(x')$. We may now suppose that, to each point x. at which $\phi(x)$ is continuous, there corresponds the point $\xi = \phi(x)$ on the ξ-segment. but that, to each point x'. at which $\phi(x)$ is discontinuous. there corresponds the closed interval $(\phi(x'-0), \phi(x'-0))$ on the ξ-segment If $G^{(x)}$ denote any set of points in the x-segment (a, b). there is a corresponding set $G^{(\xi)}$ in the ξ-segment; to any point of $G^{(x)}$. at which $\phi(x)$ is discontinuous. there corresponds a component of $G^{(\xi)}$. consisting of all the points of a closed interval. If $G^{(x)}$ consists of all the points of a closed interval. $G^{(\xi)}$ consists of all the points of a closed interval. Similarly. if $G^{(x)}$ consists of the points of an open interval, $G^{(\xi)}$ consists also of the points of an open interval. It follows that, if $G^{(x)}$ is an open set. consisting necessarily of a set of open intervals, the corresponding set $G^{(\xi)}$ is open. Consequently. if $G^{(x)}$ is closed, so also is $G^{(\xi)}$. It now follows that. if $G^{(x)}$ be measurable (B), on the x-segment, $G^{(\xi)}$ is measurable (B). on the ξ-segment. If $G^{(x)}$ be any measurable set, it is contained in a set $G_1^{(x)}$. and it contains a set $G_2^{(x)}$, such that $G_1^{(x)}$ and $G_2^{(x)}$ are measurable (B). and that the set $G_1^{(x)} - G_2^{(x)}$ has the measure zero. The two sets $G_1^{(\xi)}$. $G_2^{(\xi)}$ are both measurable (B), but it is not necessarily the case that $G_1^{(\xi)} - G_2^{(\xi)}$ has measure zero. Unless $G_1^{(x)}$ and $G_2^{(x)}$ can be so chosen that the measure of $G_1^{(\xi)} - G_2^{(\xi)}$ is zero, the set $G^{(\xi)}$ is not necessarily measurable. Even if the function $\phi(x)$ be continuous in (a, b), it is not necessarily the case that there corresponds a measurable set $G^{(\xi)}$ to any measurable set $G^{(x)}$. Only when $\phi(x)$ is such that, to every x-set of measure zero, there corresponds a ξ-set of measure zero, is $G^{(\xi)}$ necessarily measurable when $G^{(x)}$ is measurable.

In general, there corresponds to a measurable set $G^{(x)}$ a set $G^{(\xi)}$, with exterior and interior measures $m_e(G^{(\xi)})$, $m_i(G^{(\xi)})$.

The exterior and interior measures of the set $G^{(\xi)}$, *on the* ξ*-segment, which corresponds to the set* G, *on the* x*-segment, are taken to define the upper and lower variations respectively, of the function* $\phi(x)$ *over the set* G*; these may be denoted by* $\overline{V}^{(G)}\phi(x)$, $\underline{V}^{(G)}\phi(x)$ *respectively. In case*

$$\overline{V}^{(G)}\phi(x) = \underline{V}^{(G)}\phi(x),$$

their common value $V^{(G)}\phi(x)$ *is said to be the variation of the monotone function* $\phi(x)$ *over the set of points* G.

It is clear from the definition that

$$\overline{V}^{(G)}\phi(x) + \underline{V}^{C(G)}\phi(x) = \phi(b) - \phi(a).$$

If $G^{(x)}$ be measurable (B), since $G^{(\xi)}$ is also measurable (B), it follows that $\phi(x)$ has a definite variation over the set $G^{(x)}$. The variation of $\phi(x)$ over a closed interval (α, β), contained in (a, b), is $\phi(\beta+0) - \phi(\alpha-0)$; and the variation over the open interval (α, β) is $\phi(\beta-0) - \phi(\alpha+0)$. In case $\alpha = a$, we take $\phi(a-0) = \phi(a)$; and in case $\beta = b$, we take $\phi(b+0) = \phi(b)$.

If the monotone function $\phi(x)$ is absolutely continuous in (a, b), the ξ-set which corresponds to an x-set of measure zero is also of measure zero. For the x-set may then be enclosed in a set of intervals (α_r, β_r) such that the sum of their measures is arbitrarily small, and thus such that the sum $\Sigma \mid \phi(\beta_r) - \phi(\alpha_r) \mid$ is arbitrarily small. Hence the corresponding ξ-set has measure zero. It follows that, to any measurable set $G^{(x)}$ on the x-segment, there corresponds a measurable set $G^{(\xi)}$ on the ξ-segment, of which the measure is the variation of $\phi(x)$ over $G^{(x)}$.

If $\phi(x)$ be a function of bounded variation, defined for the interval (a, b), $\phi(x) - \phi(a)$ is expressible as the difference of two bounded monotone non-diminishing functions $\phi_1(x)$ and $\phi_2(x)$, where $\phi_1(x)$ and $-\phi_2(x)$ are the positive and the negative variation of $\phi(x)$ in (a, x), (see § 244).

The sum of the variations of $\phi_1(x)$, $\phi_2(x)$ *over a set* $G^{(x)}$, *for which these variations exist, may be defined to be the total variation of the function* $\phi(x)$, *of bounded variation, over the set* $G^{(x)}$.

If the absolutely continuous function $\phi(x)$ be not monotone, since it is of bounded variation (see § 243[1]) it is the difference $\phi_1(x) - \phi_2(x)$ of two monotone functions, each of which is absolutely continuous. Therefore the variation of $\phi(x)$ over any x-set of measure zero is zero. If e be any measurable set of points in (a, b), the total variation of $\phi(x)$ over e is the sum of the variations of $\phi_1(x)$ and $\phi_2(x)$ over e; and it may be regarded as a function $\Phi(e)$, of the measurable set e. This function $\Phi(e)$ is completely additive, in virtue of the property of the measures of measurable sets of points.

FUNCTIONS OF TWO VARIABLES THAT ARE OF BOUNDED VARIATION

253. The definition of a function of bounded variation, in an interval, can be extended to the case of a function of two or more variables, defined in a given cell. It will be sufficient to consider the case of a function $f(x^{(1)}, x^{(2)})$, defined in a closed cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. Let a set of $n-1$ points, in the cell, two of which are the points $(a^{(1)}, a^{(2)})$, $(b^{(1)}, b^{(2)})$, be considered; they may be denoted by

$$(x_0^{(1)}, x_0^{(2)}),\ (x_1^{(1)}, x_1^{(2)}),\ \dots\ (x_r^{(1)}, x_r^{(2)}),\ \dots\ (x_n^{(1)}, x_n^{(2)});$$

where $\quad x_0^{(1)} = a^{(1)}, \quad x_0^{(2)} = a^{(2)}, \quad x_n^{(1)} = b^{(1)}, \quad x_n^{(2)} = b^{(2)}.$

Further, let the points be such that, for each value of r $(= 1, 2, 3, \dots n)$,

$$x_r^{(1)} \geqq x_{r-1}^{(1)}, \quad x_r^{(2)} \geqq x_{r-1}^{(2)};$$

and consider the sum

$$\sum_{r=1}^{r=n} | f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) |.$$

If this sum does not exceed some fixed positive number, whatever value n may have, and however the points are chosen, subject to the conditions above stated, the function $f(x^{(1)}, x^{(2)})$ is said* to have bounded variation in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. The upper boundary of the above sum may then be denoted by $V_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})$.

A function $\phi(x^{(1)}, x^{(2)})$ is said to be a monotone non-diminishing function in the cell, if $\phi(x^{(1)}, x^{(2)}) \geqq \phi(x^{(1)\prime}, x^{(2)\prime})$ for every pair of points $(x^{(1)}, x^{(2)})$, $(x^{(1)\prime}, x^{(2)\prime})$ such that $x^{(1)} \geqq x^{(1)\prime}$, $x^{(2)} \geqq x^{(2)\prime}$. If

$$\phi(x^{(1)}, x^{(2)}) \leqq \phi(x^{(1)\prime}, x^{(2)\prime}),$$

for $x^{(1)} \geqq x^{(1)\prime}$, $x^{(2)} \geqq x^{(2)\prime}$, we have the definition of a non-increasing monotone function.

If we consider separately those terms of the sum

$$\sum_{r=1}^{r=n} \{ f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) \}$$

which are positive and those which are negative, and denote the two parts respectively by Σ_1, Σ_2, we have

$$f(b^{(1)}, b^{(2)}) - f(a^{(1)}, a^{(2)}) = \Sigma_1 \{ f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) \} \\ + \Sigma_2 \{ f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) \};$$

further

$$\Sigma_1 \{ f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) \}, \ - \Sigma_2 \{ f(x_r^{(1)}, x_r^{(2)}) - f(x_{r-1}^{(1)}, x_{r-1}^{(2)}) \}$$

are such that their sum cannot exceed $V_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})$.

It follows that Σ_1 and $-\Sigma_2$ have upper boundaries

$$\tfrac{1}{2} [V_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)}) + f(b^{(1)}, b^{(2)}) - f(a^{(1)}, a^{(2)})],$$

and $\quad \tfrac{1}{2} [V_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)}) - f(b^{(1)}, b^{(2)}) + f(a^{(1)}, a^{(2)})],$

---

* See Arzelà, *Bologna Rend.* vol. IX (1904–5).

which may be denoted respectively by

$$P_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)}), \text{ and } N_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)}).$$

These numbers may be called the *total positive variation* and the *total negative variation of* $f(x^{(1)}, x^{(2)})$ in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

If we take any point $(x^{(1)}, x^{(2)})$ in the given cell, we may denote by

$$P(x^{(1)}, x^{(2)}),\quad N(x^{(1)}, x^{(2)})$$

the positive and the negative total variation of the function in the cell $(a^{(1)}, a^{(2)}; x^{(1)}, x^{(2)})$. We have then

$$V_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} f(x^{(1)}, x^{(2)}) = P(x^{(1)}, x^{(2)}) + N(x^{(1)}, x^{(2)}),$$

$$f(x^{(1)}, x^{(2)}) - f(a^{(1)}, a^{(2)}) = P(x^{(1)}, x^{(2)}) - N(x^{(1)}, x^{(2)}),$$

and therefore, since the two functions $P(x^{(1)}, x^{(2)})$, $N(x^{(1)}, x^{(2)})$ are clearly monotone and non-diminishing, we see that $f(x^{(1)}, x^{(2)})$ can be expressed as the difference of two such functions, or also as the difference of two monotone non-increasing functions.

It is seen at once that any function which is the difference of two bounded monotone functions is of bounded variation in the cell. It has thus been shewn that:

The necessary and sufficient condition that a function $f(x^{(1)}, x^{(2)})$ *should be of bounded variation in the cell* (a, b) *is that it should be expressible as the difference of two bounded monotone functions.*

If U_r, L_r are the upper and lower boundaries of $f(x^{(1)}, x^{(2)})$ in the cell $(x_{r-1}^{(1)}, x_{r-1}^{(2)}; x_r^{(1)}, x_r^{(2)})$, we consider the sum $\sum_{r=1}^{r=n}(U_r - L_r)$, where the $n+1$ points satisfy the same conditions as before. If this sum is less than a fixed positive number, however such a system of points be chosen, the upper boundary of the sum is called the total fluctuation of the function in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, and it may be denoted by $F_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})$. The function is then said to be of bounded total fluctuation in the cell.

We see that

$$f(x^{(1)} + h^{(1)}, x^{(2)} + h^{(2)}) - f(x^{(1)}, x^{(2)}) \leq F_{(a^{(1)},\, a^{(2)})}^{(x^{(1)}+h^{(1)},\, x^{(2)}+h^{(2)})} - F_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})},$$

and

$$\geq F_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} - F_{(a^{(1)},\, a^{(2)})}^{(x^{(1)}+h^{(1)},\, x^{(2)}+h^{(2)})},$$

and that the function $F_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} f(x^{(1)}, x^{(2)})$ is monotone and non-diminishing. As in § 248, it thus appears that the necessary and sufficient condition that $f(x^{(1)}, x^{(2)})$ should be of bounded total fluctuation is that it should be the difference of two bounded monotone functions. Comparing this with the corresponding theorem for functions of bounded variation, we see that the classes of functions of bounded variation, and of bounded total fluctuation, are identical.

254. A definition of functions of bounded variation, differing from that given in § 253, has been employed by Hardy[*] and by Krause[†].

Let us consider the number

$$\Delta_{(\alpha^{(1)},\,\alpha^{(2)})}^{(\beta^{(1)},\,\beta^{(2)})} f(x^{(1)}, x^{(2)}) \equiv f(\beta^{(1)}, \beta^{(2)}) - f(\beta^{(1)}, \alpha^{(2)}) - f(\alpha^{(1)}, \beta^{(2)}) - f(\alpha^{(1)}, \alpha^{(2)}),$$

related to the cell $(\alpha^{(1)}, \alpha^{(2)}: \beta^{(1)}, \beta^{(2)})$, where $\beta^{(1)} \geq \alpha^{(1)}$, $\beta^{(2)} \geq \alpha^{(2)}$. It is seen at once that, if a net be fitted on to the cell $(a^{(1)}, a^{(2)}, b^{(1)}, b^{(2)})$,

$$\Delta_{(a^{(1)},\,a^{(2)})}^{(b^{(1)},\,b^{(2)})} f(x^{(1)}, x^{(2)}) = \Sigma\Delta_{(\alpha^{(1)},\,\alpha^{(2)})}^{(\beta^{(1)},\,\beta^{(2)})} f(x^{(1)}, x^{(2)});$$

where the summation on the right-hand side is taken for all the meshes $(\alpha^{(1)}, \alpha^{(2)}; \beta^{(1)}, \beta^{(2)})$ which make up the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

If the sum $\Sigma \mid \Delta_{(\alpha^{(1)},\,\alpha^{(2)})}^{(\beta^{(1)},\,\beta^{(2)})} f(x^{(1)}, x^{(2)}) \mid$ is less than some fixed positive number, for all possible nets that can be fitted on to the cell $(a^{(1)}, a^{(2)}: b^{(1)}, b^{(2)})$, and if further the function $f(x^{(1)}, x^{(2)})$ is, for each value of $x^{(1)}$, a function of bounded variation with respect to $x^{(2)}$, and for each value of $x^{(2)}$ a function of bounded variation with respect to $x^{(1)}$, then $f(x^{(1)}, x^{(2)})$ is said to be a function of bounded variation in the cell.

It has been pointed out[‡] by W. H. Young that there is a certain redundancy in the last two conditions. It will in fact be shewn that it is sufficient that $f(x^{(1)}, x^{(2)})$ should be of bounded variation with respect to $x^{(2)}$, for one fixed value of $x^{(1)}$, and with respect to $x^{(1)}$ for one fixed value of $x^{(2)}$: it being assumed that the first condition is satisfied.

If the first condition is satisfied, let $V_{(a^{(1)},\,a^{(2)})}^{(b^{(1)},\,b^{(2)})} f(x^{(1)}, x^{(2)})$ be the upper boundary of the sum $\Sigma \mid \Delta_{(\alpha^{(1)},\,\alpha^{(2)})}^{(\beta^{(1)},\,\beta^{(2)})} f(x^{(1)}, x^{(2)}) \mid$ for all nets that can be fitted on to the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. If the sum Σ be divided into two parts, Σ_1 and Σ_2, where Σ_1 denotes the sum of those terms for which Δ is positive, and Σ_2 denotes the sum of those terms for which Δ is negative, we see that $\Sigma_1 \Delta f(x^{(1)}, x^{(2)})$ and $-\Sigma_2 \Delta f(x^{(1)}, x^{(2)})$ have finite upper boundaries

$$P_{(a^{(1)},\,a^{(2)})}^{(b^{(1)},\,b^{(2)})} f(x^{(1)}, x^{(2)}), \quad N_{(a^{(1)},\,a^{(2)})}^{(b^{(1)},\,b^{(2)})} f(x^{(1)}, x^{(2)}).$$

The two functions $P(x^{(1)}, x^{(2)})$, $N(x^{(1)}, x^{(2)})$, defined as

$$P_{(a^{(1)},\,a^{(2)})}^{(x^{(1)},\,x^{(2)})} f(x^{(1)}, x^{(2)}), \quad N_{(a^{(1)},\,a^{(2)})}^{(x^{(1)},\,x^{(2)})} f(x^{(1)}, x^{(2)})$$

respectively, are monotone functions, in the sense that, if $x^{(1)} \leq \bar{x}^{(1)}$, $x^{(2)} \leq \bar{x}^{(2)}$, then $P(x^{(1)}, x^{(2)}) \leq P(\bar{x}^{(1)}, \bar{x}^{(2)})$, $N(x^{(1)}, x^{(2)}) \leq N(\bar{x}^{(1)}, \bar{x}^{(2)})$.

We have

$$f(x^{(1)}, x^{(2)}) = -f(a^{(1)}, a^{(2)}) + f(x^{(1)}, a^{(2)}) + f(a^{(1)}, x^{(2)}) - P(x^{(1)}, x^{(2)}) \\ - N(x^{(1)}, x^{(2)}).$$

Assuming that $f(x^{(1)}, a^{(2)})$ is a function of $x^{(1)}$, of bounded variation in the interval $(a^{(1)}, b^{(1)})$, it is equal to $p(x^{(1)}) - n(x^{(1)})$, where $p(x^{(1)})$, $n(x^{(1)})$ are monotone functions of $x^{(1)}$. Similarly, assuming that $f(a^{(1)}, x^{(2)})$ is of bounded variation in the interval $(a^{(2)}, b^{(2)})$, of $x^{(2)}$, it is representable as the difference $p'(x^{(2)}) - n'(x^{(2)})$ of two monotone functions of $x^{(2)}$.

---

[*] *Quarterly Journal of Math.* vol. XXXVII (1906), p. 57 *et seq.*

[†] *Leipziger Ber.* vol. LV (1903), pp. 164, 239; see also Vergerio, *Gior. di Mat.* vol. XLIX (1911), p. 181.

[‡] *Proc. Lond. Math. Soc.* (2), vol. XI (1913), p. 142.

If
$$\bar{P}(x^{(1)}, x^{(2)}) = P(x^{(1)}, x^{(2)}) + p(x^{(1)}) + p'(x^{(2)}),$$
$$\bar{N}(x^{(1)}, x^{(2)}) = N(x^{(1)}, x^{(2)}) + n(x^{(1)}) + n'(x^{(2)}),$$
we have
$$f(x^{(1)}, x^{(2)}) = \bar{P}(x^{(1)}, x^{(2)}) - \bar{N}(x^{(1)}, x^{(2)}) - f(a^{(1)}, a^{(2)}),$$
where $\bar{P}$, $\bar{N}$ are monotone non-diminishing functions. It has now been shewn that a function $f(x^{(1)}, x^{(2)})$ of bounded variation, in accordance with the definition here given, is expressible as the difference of two bounded monotone functions. It is therefore of bounded variation, according to the definition in § 253.

The definition here given of a function of bounded variation is less general than that of Arzelà, given in § 253. An example has been constructed* by Küstermann, of a function which is of bounded variation in accordance with Arzelà's definition, but not in accordance with that here considered.

The variation of $f(x^{(1)}, k^{(2)})$ for any fixed value $k^{(2)}$, of $x^{(2)}$, clearly does not exceed the sum of the variations of $f(x^{(1)}, a^{(2)})$, of $P(x^{(1)}, k^{(2)})$, and of $N(x^{(1)}, k^{(2)})$. It is therefore bounded, on the assumption that the variation of $f(x^{(1)}, a^{(2)})$ is bounded. Similarly, on the assumption that $f(a^{(1)}, x^{(2)})$ is bounded, that of $f(k^{(1)}, x^{(2)})$ is bounded, for any fixed value $k^{(1)}$, of $x^{(1)}$. In each case the converse holds good.

If a function $\phi(x^{(1)}, x^{(2)})$, defined in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ be such that, for any finite, or infinite, set of non-overlapping cells (any two of which may however have portions of their boundaries in common), the sum
$$\sum_{r=1} \left| \Delta_{(\alpha_r{}^{(1)}, \alpha_r{}^{(2)})}^{(\beta_r{}^{(1)}, \beta_r{}^{(2)})} \phi(x^{(1)}, x^{(2)}) \right|$$
is less than an arbitrarily fixed positive number η, for all sets of such cells which satisfy the condition that the sum of their measures is less than some positive number ϵ_η, the function $\phi(x^{(1)}, x^{(2)})$ is then said to be *absolutely continuous in the cell* $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. The proofs of the theorems given in § 243¹, for the case of linear sets, can be applied, if two-dimensional nets are employed instead of linear nets, to prove that an absolutely continuous function $\phi(x^{(1)}, x^{(2)})$ is of bounded variation, in accordance with the definition of Hardy and Krause, provided $\phi(x^{(1)}, a^{(2)})$, $\phi(a^{(1)}, x^{(2)})$ are of bounded variation. It may be remarked that the whole theory given here is applicable by extension to the case of p-dimensions, where $\Delta_{(\alpha^{(1)}, \alpha^{(2)}, \ldots \alpha^{(p)})}^{(\beta^{(1)}, \beta^{(2)}, \ldots \beta^{(p)})} f(x^{(1)}, x^{(2)}, \ldots x^{(p)})$ is defined as
$$\Delta_{(\alpha^{(p)})}^{(\beta^{(p)})} \cdot \Delta_{(\alpha^{(1)}, \alpha^{(2)} \ldots \alpha^{(p-1)})}^{(\beta^{(1)}, \beta^{(2)} \ldots \beta^{(p-1)})} f(x^{(1)}, x^{(2)}, \ldots x^{(p)}).$$

---

* *Math. Annalen*, vol. LXXVII (1916), p. 474.

Thus, for example,

$$\begin{aligned}\Delta_{(\alpha^{(1)},\,\alpha^{(2)},\,\alpha^{(3)})}^{(\beta^{(1)},\,\beta^{(2)},\,\beta^{(3)})} f(x^{(1)}, x^{(2)}, x^{(3)}) &= \Delta_{\alpha^{(3)}}^{\beta^{(3)}} \cdot \Delta_{(\alpha^{(1)},\,\alpha^{(2)})}^{(\beta^{(1)},\,\beta^{(2)})} f(x^{(1)}, x^{(2)}, x^{(3)})\\ &= [f(\beta^{(1)}, \beta^{(2)}, \beta^{(3)}) - f(\beta^{(1)}, \alpha^{(2)}, \beta^{(3)}) - f(\alpha^{(1)}, \beta^{(2)}, \beta^{(3)})\\ &\qquad\qquad - f(\alpha^{(1)}, \alpha^{(2)}, \beta^{(3)})]\\ &\quad - [f(\beta^{(1)}, \beta^{(2)}, \alpha^{(3)}) - f(\beta^{(1)}, \alpha^{(2)}, \alpha^{(3)}) - f(\alpha^{(1)}, \beta^{(2)}, \alpha^{(3)})\\ &\qquad\qquad - f(\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)})].\end{aligned}$$

QUASI-MONOTONE FUNCTIONS

255. A function $f(x^{(1)}, x^{(2)})$ may satisfy the condition

$$\Delta_{(x^{(1)},\,x^{(2)})}^{(\bar{x}^{(1)},\,\bar{x}^{(2)})} f(x^{(1)}, x^{(2)}) \geqq 0, \text{ for } \bar{x}^{(1)} \geqq x^{(1)},\ \bar{x}^{(2)} \geqq x^{(2)},$$

and yet $f(x^{(1)}, x^{(2)})$ need not be monotone and non-diminishing with respect to $x^{(2)}$, for each constant value of $x^{(1)}$; nor need it be monotone and non-diminishing with respect to $x^{(1)}$, for each constant value of $x^{(2)}$. For the expression may be written in either of the forms

$$\{f(\bar{x}^{(1)}, \bar{x}^{(2)}) - f(x^{(1)}, \bar{x}^{(2)})\} - \{f(\bar{x}^{(1)}, x^{(2)}) - f(x^{(1)}, x^{(2)})\},$$

$$\{f(\bar{x}^{(1)}, \bar{x}^{(2)}) - f(\bar{x}^{(1)}, x^{(2)})\} - \{f(x^{(1)}, \bar{x}^{(2)}) - f(x^{(1)}, x^{(2)})\}.$$

The expression in any one of the four brackets may have either sign, consistently with the condition stated, except that the first bracket in either case cannot be negative whilst the second is positive.

We may consequently consider four types of functions $f(x^{(1)}, x^{(2)})$, which satisfy the condition $\Delta_{(x^{(1)},\,x^{(2)})}^{(\bar{x}^{(1)},\,\bar{x}^{(2)})} f(x^{(1)}, x^{(2)}) \geqq 0$, for $\bar{x}^{(1)} \geqq x^{(1)}$, $\bar{x}^{(2)} \geqq x^{(2)}$, viz.:

(1). Those for which $f(x^{(1)}, x^{(2)})$ is monotone and non-diminishing with respect to $x^{(1)}$, for every constant value of $x^{(2)}$; and is also monotone and non-diminishing with respect to $x^{(2)}$, for every constant value of $x^{(1)}$.

(2). Those for which $f(x^{(1)}, x^{(2)})$ is monotone and non-diminishing with respect to $x^{(1)}$, for every constant value of $x^{(2)}$, and is monotone and non-increasing with respect to $x^{(2)}$, for every constant value of $x^{(1)}$.

(3). Those for which $f(x^{(1)}, x^{(2)})$ is monotone and non-increasing with respect to $x^{(1)}$, and is monotone and non-diminishing with respect to $x^{(2)}$.

(4). Those for which $f(x^{(1)}, x^{(2)})$ is monotone and non-increasing with respect to $x^{(1)}$, and is monotone and non-increasing with respect to $x^{(2)}$.

A function $f(x^{(1)}, x^{(2)})$ which satisfies the condition $\Delta f(x^{(1)}, x^{(2)}) \geqq 0$, for every pair of points, or else the condition $\Delta f(x^{(1)}, x^{(2)}) \leqq 0$, for every pair of points, and also one of the four conditions given above, everywhere in its domain, may be said to be a *quasi-monotone function* of one of the four types which correspond to the conditions (1), (2), (3), (4), respectively.

It is easy to see that this definition can be extended* to the case of functions of any number of variables.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1917), p. 277. The term "monotone" is there applied to a function described above as quasi-monotone.

Since

$$f(x^{(1)}, x^{(2)}) = \Delta f(x^{(1)}, x^{(2)}) + f(x^{(1)}, a^{(2)}) + f(a^{(1)}, x^{(2)}) - f(a^{(1)}, a^{(2)})$$

it is seen that every quasi-monotone function is the sum of a monotone function of the two variables, and of monotone functions of the separate variables, such that any one of the three may be either non-increasing or non-diminishing. The converse does not however hold; in fact a function can be monotone without necessarily being quasi-monotone.

For example, let the monotone function $f(x^{(1)}, x^{(2)})$ be defined in the cell $(-1, -1; 1, 1)$ as $(x^{(1)} + 1)(x^{(2)} + 1)$, or $(x^{(1)} + 2)(x^{(2)} + 2)$, according as $x^{(1)} + x^{(2)}$ is < 0, or $\geqq 0$. It is then seen that $\Delta f(x^{(1)}, x^{(2)})$ is positive over $(0, 0; 1, 1)$, and is negative over $(0, -\frac{1}{2}; 1, 1)$; thus the function is not quasi-monotone.

THE MAXIMA, MINIMA, AND LINES OF INVARIABILITY OF CONTINUOUS FUNCTIONS

256. Consider a point x_1 within the interval (a, b), in which a continuous function is defined, it may happen that a neighbourhood $(x_1 - \delta, x_1 + \delta)$ of the point x_1 can be found, by taking δ sufficiently small, which is such that $f(x)$ has the same value at all points in the neighbourhood; then the point x_1 is called a *point of linear invariability* of the function. If the same holds for a neighbourhood of x_1 on the right only, or on the left only, then the point x_1 is called a *limiting point of linear invariability.*

It can be shewn that, if a point x_1 of linear invariability exist, and the function be not constant in the whole interval (a, b), then there exist two limiting points of linear invariability, one of which, however, may be at one of the ends of the interval (a, b). Suppose the function not to be constant throughout the interval (x_1, b); the points x of this interval may be divided into two classes, in one of which x is such that, in the interval (x_1, x), the function has the constant value $f(x_1)$, and in the other class x is such that (x_1, x) contains points at which the function has values differing from $f(x_1)$; a section is thus made of the interval (x_1, b). This section defines a point which is the required limiting point of the linear invariability. If the same argument be applied to the interval (a, x_1) we see that there is another limiting point in this interval, unless the function be throughout equal to $f(x_1)$.

In the interval (a, b) there may be a finite set, or an indefinitely great, but enumerable, set of lines of invariability, each point within such a line is a point of invariability, and the ends of such lines are limiting points of invariability.

If the point x_1 be not a point of invariability, it may happen that a neighbourhood $(x_1 - \epsilon, x_1 + \epsilon')$ exists, such that, for every point in the interior of this neighbourhood, not identical with x_1, the condition $f(x) < f(x_1)$ is satisfied; in this case x_1 is said to be a point at which the function has a *proper maximum*. In case the neighbourhood be such that at every point x within it, except at x_1, the condition $f(x) > f(x_1)$ is satisfied, the point x_1 is said to be a point at which the function has a *proper minimum*.

It may happen that, when x_1 is not a proper maximum, a neighbourhood $(x_1 - \epsilon, x_1 + \epsilon')$ exists which is such that at no point within it the condition $f(x) > f(x_1)$ is satisfied, whilst not at every point is the condition $f(x) < f(x_1)$ satisfied. If this is the case for every neighbourhood interior to $(x_1 - \epsilon, x_1 + \epsilon')$ then x_1 is said to be a point at which there is an *improper maximum* of the function. If the condition $f(x) \geqq f(x_1)$ is satisfied, but the condition $f(x) > f(x_1)$ is not everywhere satisfied, then x_1 is said to be a point at which there is an *improper minimum* of the function.

A line of invariability, of which the end-points are α, β, these being both interior to (a, b), is said to be a maximum of the function, if both α, β be improper maxima, and it is said to be a minimum, if both α, β be improper minima.

It is clear that, in any arbitrarily small neighbourhood of an improper maximum or minimum, there is an indefinitely great number of points at which the functional value is equal to that at the maximum or minimum.

At any maximum or minimum there is a greatest neighbourhood $(x_1 - \delta, x_1 + \delta')$ at every interior point of which the condition

$$f(x) < f(x_1),\ f(x) \leqq f(x_1),\ \text{or}\ f(x) > f(x_1),\ f(x) \geqq f(x_1)$$

is satisfied. At end-points of such greatest neighbourhood, it follows from the condition of continuity of the function, that the functional value is equal to $f(x_1)$, unless such end-point coincides with a or with b

It has been shewn in § 213 that there exists either one point, or a set of points, in (a, b), such that the functional value at this point, or at all the points of the set, is greater than at all other points in the interval; and it is to be remarked that this set of points may contain lines of invariability. Every such point, unless it be an end-point, is said to be a point of *absolute maximum* of the function in the interval (a, b), and may be either a proper or an improper maximum. A similar definition applies to an *absolute minimum*.

In case an extreme point of the continuous function (see § 209) be at a, or at b, such point is spoken of as an upper or lower extreme, but not always as a maximum or minimum of the function. If $f(a)$ and $f(b)$ be equal, and the function be not constant in (a, b), then there is at least one maximum or one minimum point, or one line of invariability, in the interior of (a, b). This is also true when $f(a) \neq f(b)$, unless the function be monotone.

257. *If, within the interval (a, b), there be two points, or two lines of invariability, at which the continuous function is a maximum, proper or improper, then there is between them at least one point, or one line of invariability, at which the function is a proper or improper minimum; thus maxima and minima occur alternately.*

Suppose that α, β are two points at which the function is a maximum, and that (α, β) is not entirely a line of invariability; also that no maximum occurs between α and β. We know that, between α and β, there is a point, or a set of points, at which the function is less than at all other points in the sub-interval; and since α and β cannot belong to such set, there is therefore a minimum at a point, or at points on a line of invariability, between α and β, and this minimum is less than either of the maxima at α and β.

Between a maximum and the next minimum of a continuous function the function is said to make an oscillation, the amplitude of which is the excess of the maximum over the minimum.

If x_1 be a point in (a, b), it may be possible to choose ϵ so small that, within the interval $(x_1, x_1 + \epsilon)$, no maxima or minima occur, so that the function is monotone in this interval. It may however be the case that, however small ϵ is taken, there still occur maxima and minima in $(x_1, x_1 + \epsilon)$. In this case the number of oscillations of the function must be indefinitely great, however small ϵ may be chosen; for if there were a finite number only, a number ϵ_1 could be found such that all the maxima and minima were in the interval $(x_1 + \epsilon_1, x_1 + \epsilon)$, and thus in $(x_1, x_1 + \epsilon_1)$ the function would be monotone, which is contrary to the hypothesis made.

It thus appears that, in the neighbourhood of a particular point, a continuous function may have an indefinitely great number of oscillations. An improper maximum or minimum, not in a line of invariability, is certainly such a point.

The proper maxima and minima of a continuous function form an enumerable, or a finite, set of points.

Consider $(x_1 - \epsilon, x_1 + \eta)$, the greatest neighbourhood of a point of proper maximum x_1, which is such that, for all other points x within the neighbourhood, $f(x) < f(x_1)$. There can, in a finite interval, be only a finite number of such points x_1 for which $\epsilon > \alpha$, $\eta > \alpha$, where α is a fixed positive number; for if there were an infinite number of such points, they would have a limiting point ξ, and we could choose two points, x_1', x_1'' of the set, such that the distance of each from ξ is less than $\frac{1}{2}\alpha$; then each of these points would lie within the neighbourhood belonging to the other, and thus we should have $f(x_1') > f(x_1'')$, and also $f(x_1'') > f(x_1')$, which is impossible; thus the set must be finite. Now choose a sequence of descending values of α which converges to zero, say $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$; the number m_n of maxima x_1 such that for each $\epsilon > \alpha_n$, $\eta > \alpha_n$ being finite, we have $m_1, m_2, \ldots m_n, \ldots$ all finite; and hence the whole set of maxima forms an enumerable set.

If x_1 be an improper maximum point, and $f(x_1) = A$, then a neighbourhood $(x_1 - \epsilon, x_1 + \eta)$ can be found which contains an infinite set of points G_A such that $f(x) = A$, for each point of the set. If x' be an isolated point of the set G_A, then x' is clearly a proper maximum of the function: and if x'' be a point of G_A, which is a limiting point of the set, x'' is an improper maximum. The points $x_1 - \epsilon$, $x_1 + \eta$ need not be maxima, even though they be limiting points of G_A. The condition of continuity of the function ensures that the set G_A is a closed one; for, at any limiting point of the set, the functional value is the limit of a sequence, each member of which is A, and this value is therefore itself A.

Corresponding to a given A, there may be a finite, or an infinite, set of detached intervals such as $(x_1 - \epsilon, x_1 + \eta)$, each one of which contains a closed set such that each isolated point of it is a proper maximum, and each limiting point (except an end-point) is an improper maximum. The set G_A may contain perfect components, and thus the improper maxima at which A is the functional value may form a set of the cardinal number of the continuum. A similar result holds for minima

It can further be shewn that the values of a continuous function at all its maxima and minima form a set which is either finite or enumerably infinite.

258. If, in the interval (a, b), the function have only a finite number of maxima and minima, counting any line of invariability which is a maximum or minimum as one maximum or minimum, the interval can be divided into a finite number of parts in each of which the function is monotone; the function is then said to be* *in general monotone* (abtheilungsweise monoton).

If the function have an indefinitely great number of maxima and minima, which occur either at points or at lines of invariability, the function then makes an infinite number of oscillations; and these may occur in the neighbourhoods either of a finite number of points, or of an infinite number of points.

It can be shewn that, in the case of a continuous function, although there may be an infinite number of oscillations of the function, there can be only a finite number of which the amplitude exceeds an arbitrarily small fixed number σ.

---

* This term is due to C. Neumann; see his work *Ueber die nach Kreis- Kugel- und Cylinderfunctionen fortschreitenden Reihen*, Leipsic, 1881.

For it has been shewn in § 217 that a number ϵ can be determined, such that, in any sub-interval of length ϵ, the fluctuation of the function does not exceed σ; therefore in each of the sub-intervals

$$(a, a+\epsilon),\ (a+\epsilon, a+2\epsilon), \ldots (a+n\epsilon, b),$$

the fluctuation of the function is not greater than σ. It follows that no oscillation of the function which is greater than σ can be completed in one of these sub-intervals, and that such an oscillation must require two at least of these sub-intervals for its completion; hence the number of such oscillations in (a, b) cannot exceed the finite number n. As the number σ is diminished indefinitely, it may happen that the number of oscillations of which the amplitude exceeds σ is increased indefinitely.

If the point a is an isolated point of discontinuity of a function $f(x)$, and if $\overline{f(a+0)}$, $\underline{f(a+0)}$ be unequal, there is an infinite number of oscillations in the interval $(a, a+\epsilon)$. This may also be the case if $f(a+0)$ have a unique value.

THE DERIVATIVES OF FUNCTIONS

259. If a function $f(x)$ be defined for all points in the interval (a, b), then, for a point x_1 in this interval, we may regard the function

$$\frac{f(x)-f(x_1)}{x-x_1}$$

as a function $F(x)$, of x, which is defined for all values of x in (a, b), except for the point x_1. This function $F(x)$, although undefined at the point x_1, has finite or infinite functional limits at that point, in accordance with the definitions in § 220.

If the limits $F(x_1+0)$, $F(x_1-0)$ both exist, and have the same value, either finite, or $+\infty$, or $-\infty$, this value is called the *differential coefficient* at x_1 of the function $f(x)$. At the point a, if $F(a+0)$ exists, it is frequently said to be the differential coefficient of $f(x)$ at a; and at the point b, if $F(b-0)$ exists, it is said to be the differential coefficient of $f(x)$ at b.

The condition that $f(x)$ may possess a finite differential coefficient at x_1 is that, corresponding to each arbitrarily chosen positive number ϵ, a neighbourhood $(x_1-\delta, x_1+\delta)$ can be found, such that

$$\left|\frac{f(\xi)-f(x_1)}{\xi-x_1}-\frac{f(\xi')-f(x_1)}{\xi'-x_1}\right|<\epsilon$$

for every pair of points ξ, ξ' which lie within this neighbourhood, or within such part of it as is interior to (a, b).

In other words, the condition is that a neighbourhood of x_1 can be found such that the fluctuation of the function $\dfrac{f(x)-f(x_1)}{x-x_1}$ within it, or within such part of it as lies in (a, b), may be as small as we please.

If, corresponding to an arbitrarily large positive number A, a neighbourhood $(x_1 - \delta, x_1 + \delta)$ can be found, such that $\frac{f(x) - f(x_1)}{x - x_1} > A$, for every point x $(\neq x_1)$ in that neighbourhood, $f(x)$ has the differential coefficient $+\infty$ at the point x_1.

When a differential coefficient of $f(x)$ exists at the point x_1, the function is said to be *differentiable* at x_1, and the differential coefficient at that point may be denoted by $f'(x_1)$.

In geometrical language, when the function $f(x)$ has a differential coefficient at x_1, the curve $y = f(x)$ is said to have a tangent at the point $(x_1, f(x_1))$. In case the differential coefficient is $+\infty$, or $-\infty$, the tangent is parallel to the y-axis, and the point is a point of inflexion of the curve.

That a function $f(x)$ may have a finite differential coefficient at x_1, it is necessary, but not sufficient, that x_1 should be a point of continuity of the function.

At a point of discontinuity x_1, of $f(x)$, there always exists a positive number σ, such that in any neighbourhood of x_1, however small, points ξ exist such that $|f(\xi) - f(x_1)| > \sigma$: hence if A be any arbitrarily great positive number, in the interval $(x_1 - \delta, x_1 - \delta)$, where $\delta < \sigma/A$, there exist points ξ such that

$$\left|\frac{f(\xi) - f(x_1)}{\xi - x_1}\right| > A,$$

and it is thus impossible that $\frac{f(x) - f(x_1)}{x - x_1}$ should have a definite finite limit at x_1. On the other hand, the condition for the existence of a finite differential coefficient, viz. that $\frac{f(x) - f(x_1)}{x - x_1}$ should have an arbitrarily small fluctuation within a sufficiently small neighbourhood of x_1, is not necessarily satisfied when the condition of continuity, viz. that $f(x)$ should have an arbitrarily small fluctuation within a sufficiently small neighbourhood of x_1, is satisfied.

If, at a point x, the function $f(x)$ have a finite differential coefficient $f'(x)$, or have an infinite differential coefficient of fixed sign, an interval $(x - \delta_\epsilon,\ x + \delta_\epsilon)$ can be so determined* that the ratio $\frac{f(x_2) - f(x_1)}{x_2 - x_1}$ is between $f'(x) + \epsilon$ and $f'(x) - \epsilon$, where x_1 is any point in the interval $(x - \delta_\epsilon, x)$, and x_2 is any point in the interval $(x, x - \delta_\epsilon)$; or that the same ratio is $> N$, an assigned positive number, if $f'(x) = -\infty$, or $< -N$, if $f'(x) = -\infty$.

A continuous function $f(x)$, defined for the interval (a, b), which has a differential coefficient at every point of the interval, is said to be *differ-*

---

* See Bromwich's *Theory of Infinite Series*, p. 490.

entiable in its domain. Continuous functions exist, which at no point in their domain possess a differential coefficient. The first example of such a function was given by Weierstrass; the construction of such functions will be considered in Vol. II.

That a continuous function possesses a differential coefficient was formerly regarded as obvious from geometrical intuition, it being supposed that such functions were necessarily representable by curves possessing definite tangents at every point. The first attempt to prove the existence of a differential coefficient of a continuous function was that of Ampère*; this proof was, however, insufficient, even in the case of those continuous functions which make only a finite number of oscillations in the intervals for which they are defined. It is now fully recognized that the class of continuous functions is much wider than that of functions capable of an approximate graphical representation; and that the conditions for the existence of definite differential coefficients are of a much more stringent character than would be the case if they were included under the bare condition of continuity of the function.

260. It may happen that, at a point x_1, the function $\frac{f(x) - f(x_1)}{x - x_1}$ possesses finite, or even indefinitely great, limits on the right and on the left, at x_1, which differ from one another; the function is then said to have *derivatives, on the right,* and *on the left,* at x_1. These are frequently spoken of as the *progressive* and *regressive differential coefficients,* or *derivatives,* respectively. A function may possess a progressive derivative and no regressive derivative, or the reverse.

When, at the point x_1, the function $f(x)$ has no differential coefficient, and neither progressive nor regressive derivatives, the function

$$\frac{f(x) - f(x_1)}{x - x_1}$$

has, in accordance with § 220, four extreme limits, an upper, and a lower one, on the right, and on the left. Any one of these four may be either finite or infinite.

There is in general an aggregate of functional limits, on the right, of $\frac{f(x) - f(x_1)}{x - x_1}$, and also an aggregate of functional limits on the left. The upper and lower limits, on the right, which are the upper and lower boundaries of the aggregate of functional limits of $\frac{f(x) - f(x_1)}{x - x_1}$ at x_1, on the right, are defined to be the *upper and lower extreme derivatives of $f(x)$, at x_1, on the right,* or simply, *the upper and lower derivatives on the right,* and these are in accordance with the notation of Scheeffer*, denoted by $D^+f(x)$, $D_+f(x)$ respectively. Similarly, the upper and lower derivatives of $f(x)$ at x_1, on the left, are the upper and lower boundaries of the aggregate of functional limits of $\frac{f(x)-f(x_1)}{x-x_1}$ on the left, and are denoted by $D^-f(x)$, $D_-f(x)$ respectively.

---

* *Journ. écol. polyt.* vol VI (1806), p. 148.

As x has the values of a sequence converging to x_1 on one side, the values of $\frac{f(x)-f(x_1)}{x-x_1}$, if convergent, may converge either to the upper or to the lower derivative of $f(x)$ at x_1, on that side, or they may converge to some number lying between these extreme derivatives. In the last case they are said to converge to a *median derivative* on that side. The upper and lower derivatives on one side, together with any median derivatives on that side which may exist, make up the aggregate of functional limits of $\frac{f(x)-f(x_1)}{x-x_1}$ at x_1, on the side under consideration.

In case the function $f(x)$ is continuous in an open interval, with x_1 as end-point, on the right of x_1, the median derivatives have every value between $D^+f(x_1)$ and $D_+f(x_1)$ (see § 229), but when $f(x)$ is not continuous in such an open interval on the right of x_1, the aggregate of derivatives on the right may be a closed discontinuous set having $D^+f(x_1)$ and $D_+f(x_1)$ for its upper and lower extremes. A similar remark applies to the case of the derivatives on the left.

It is frequently convenient in this general case to speak of the derivatives of $f(x)$ on the right, and on the left, as existent, but indefinite in value: and in this case $D^+f(x_1)$, $D_+f(x_1)$ are regarded as the limits of indeterminancy of the derivative on the right, and $D^-f(x_1)$, $D_-f(x_1)$, as those of the derivative on the left.

The definitions which have been given for the case in which the domain of the function is continuous are applicable, without essential change, to the case in which the domain is any perfect set of points. At a point of the set which is a limiting point on both sides, there exist in general the four extreme derivatives $D^+f(x)$, $D_+f(x)$, $D^-f(x)$, $D_-f(x)$, two or more of which may have equal values; and at a point of the perfect set, which is a limiting point on one side only, there exist of course only the two derivatives on that side. If the domain be any closed set, the derivatives exist only at those points which are limiting points of the set.

A function defined for a perfect set may, by the method of correspondence, be correlated with a function defined for a continuous interval, the relative order of two points in the continuous interval being the same as

---

[*] *Acta Mathematica*, vol. v (1884), p. 52. The same limits were considered by Du Bois-Reymond, *Programm*, Freiburg, 1870, also *Münch. Abh.* vol. XII (1875), p. 125, under the name Unbestimmtheitsgrenzen.

that of the corresponding points in the perfect set (see § 163); and thus all properties of derivatives of functions defined for a continuous interval have their analogues in the case in which the domain is any perfect set.

EXAMPLES

1. If $f(x) = x \sin \frac{1}{x}$, $f(0) = 0$; we have $\frac{f(h) - f(0)}{h} = \sin \frac{1}{h}$, and for arbitrarily small values of h, this oscillates between 1 and -1. The function $f(x)$, although continuous at $x = 0$, possesses no differential coefficient at that point; in fact

$$D^+ f(0) = 1,\ D_+ f(0) = -1,\ D^- f(0) = 1,\ D_- f(0) = -1.$$

2. If $f(x) = x^2 \sin \frac{1}{x}$, $f(0) = 0$, the differential coefficient $f'(x)$ exists for every value of x, and is finite. At the point $x = 0$, $f'(x)$ is zero, but has a discontinuity of the second kind.

3. Let* $f(x) = \sqrt{x}\left(1 + x \sin \frac{1}{x}\right)$, for $x > 0$; $f(x) = -\sqrt{-x}\left(1 + x \sin \frac{1}{x}\right)$, for $x < 0$; and $f(0) = 0$. In this case $f'(x)$ everywhere exists; its value at $x = 0$ is $+\infty$, and although it has a finite value at every point except at $x = 0$, it oscillates in the neighbourhood of that point between indefinitely great positive and negative values.

4.† The function defined by $f(x) = x\{1 + \frac{1}{3} \sin(\log x^2)\}$, and $f(0) = 0$, is everywhere continuous, and is monotone, but has no differential coefficient at $x = 0$.

5.‡ Let $f(x) = e^{-\frac{1}{x^2}} \sin \frac{1}{x}$, $f(0) = 0$; this function has at every point a differential coefficient, and this is continuous at $x = 0$. The differential coefficient vanishes at $x = 0$, and at an infinite number of points in the neighbourhood of $x = 0$. The function $f'(x)$, like $f(x)$, has an infinite number of oscillations in a neighbourhood of $x = 0$.

THE DIFFERENTIAL COEFFICIENTS OF CONTINUOUS FUNCTIONS

261. Let us suppose that a continuous function, defined for a continuous closed interval, is such that, at every point interior to an interval (α, β), there exists a differential coefficient; this differential coefficient may at any point have a finite value, which may be zero, or it may have an infinite value, of which, however, the sign is definite. It will be observed that $f(x)$ is assumed to be continuous at the points α, β, but it is not assumed that definite derivatives exist at those points. It will be shewn that, *unless the function be constant throughout* (α, β), *there exists at least one point in the interior of* (α, β) *at which the differential coefficient has a definite value, different from zero.*

Suppose $f(\alpha)$, $f(\beta)$ to be unequal. If they be not unequal, and the function be not constant throughout (α, β), we can replace the interval

---

* Dini-Lüroth, *Grundlagen*, p. 112.

† Pringsheim, *Encyclopädie der math. Wissensch.* II A. I, p. 22.

‡ Dini-Lüroth, *Grundlagen*, p. 313.

(α, β), by another one contained in it, for which the functional values at the ends are unequal. Let us consider the function

$$F(x) = f(x) - f(\alpha) - \frac{x - \alpha}{\beta - \alpha}\{f(\beta) - f(\alpha)\}.$$

$F(\alpha)$ and $F(\beta)$ vanish, and $F(x)$ is continuous in (α, β), and has a differential coefficient in the ordinary sense at each point, with the possible exception of α and β; therefore it follows, by the theorem of § 213, that there is at least one point x_1, in the interior of (α, β), at which $F(x)$ is a maximum or minimum: this is the case even if $F(x)$ be everywhere zero in the interval. A number ϵ can therefore be found such that

$$F(x_1 + \delta) - F(x_1),\ F(x_1 - \delta) - F(x_1)$$

have the same sign, or else vanish, provided $\delta < \epsilon$: and consequently the derivatives at x_1, on the right and left, must have opposite signs, unless both of them be zero; therefore the differential coefficient at x_1, which must exist, must be zero. It follows that $f'(x_1) - \frac{f(\beta) - f(\alpha)}{\beta - \alpha} = 0$, and thus the point x_1 is the point of which the existence was to be proved. From this theorem we deduce the following general theorem:

If $f(x)$ be continuous in the closed interval (a, b), and be such that it has a differential coefficient at every point in the interior of the interval, and if there be in (a, b) no lines of invariability, then there exists in (a, b) an everywhere dense set of points at which the differential coefficient has finite values differing from zero.

This is proved at once by applying the foregoing theorem to any interval contained in (a, b). There may be, in (a, b), infinite sets of points at which the differential coefficient is either zero or infinite.

262. The following theorem, known as the *mean value theorem* of the Differential Calculus, has been established by the proof in § 261:

If the function $f(x)$ be continuous in the closed interval (a, b), and if the differential coefficient $f'(x)$ exist at every point interior to the interval (a, b), being either finite, or infinite with fixed sign, then a point $a + \theta(b - a)$ exists, where θ is some number such that $0 < \theta < 1$, for which

$$f(b) = f(a) + (b - a) f'(a + \theta\overline{b - a}).$$

If we change the notation, so that $(x, x + h)$ is the given interval, the result may be written

$$f(x + h) = f(x) + hf'(x + \theta h).$$

It will be observed that no assumption is made as to the existence of a derivative on the right at the point a, or x, or as to the existence of a derivative on the left at b, or $x + h$. It has only been assumed that the function is continuous on the right, and on the left, at these points respectively.

In case however it is known that the limit of $f'(x)$, as $x \sim a$, has a definite value λ, then $f(x)$ has a definite derivative on the right at the point a, and its value is λ.

For, since $\dfrac{f(x)-f(a)}{x-a} = f'(a + \theta\overline{x-a})$, and the expression on the right-hand side converges to λ, as $x \sim a$, it follows that

$$\lim_{x \sim a} \frac{f(x)-f(a)}{x-a} = \lambda.$$

A corollary from the mean value theorem is that, if $f(x+h) = f(x)$, then $f'(x)$ must be zero at one point at least in the interior of the interval $(x, x+h)$.

The following theorem may be proved by means of the mean value theorem:

If $f(x)$ have a continuous differential coefficient $f'(x)$, at every point of an open interval (a, b), then $f(x+h) - f(x) = h\{f'(x) + \rho(x, h)\}$, for every pair of points x, $x+h$ of the open interval; where $\rho(x, h)$ converges to zero, as $h \sim 0$, uniformly for all points x of any closed interval (α, β) contained in (a, b).

Since $f(x+h) - f(x) = hf'(x+\theta h)$, where $0 < \theta < 1$, for any pair of points x, $x+h$, within (a, b), and since $f'(x)$ is continuous in the closed interval $(\alpha - h_1, \beta + h_1)$, where (α, β) is a closed interval contained in (a, b), and h_1 is so small that the closed interval $(\alpha - h_1, \beta + h_1)$ is also in (a, b), we have, for every value of x in (α, β), $|f'(x+\theta h) - f'(x)| < \epsilon$, if $|h| < h_1$, provided h_1 is taken sufficiently small. This follows from the uniform continuity of $f'(x)$ in the closed interval $(\alpha - h_1, \beta + h_1)$. We may therefore take $f'(x+\theta h) = f'(x) + \rho(x, h)$; where $|\rho(x, h)| < \epsilon$, provided $|h| < h_1$, and x is any point in (α, β). Since ϵ is arbitrary, $\rho(x, h)$ converges uniformly to zero, as $h \sim 0$, for all points x in (α, β).

263. An important extension of the mean value theorem is the following:

If $f(x)$ be continuous in the closed interval $(x, x+h)$, and have a differential coefficient at every point of the interval, with the possible exception of the end-points; and if $F(x)$ be another function which is also continuous in the same interval, and at every interior point has a differential coefficient, whilst at the end-points there may be no definite derivatives, or they may be zero, or infinite, then, provided $f'(x)$ and $F'(x)$ have no common zeros or common infinities in the open interval,

$$\frac{f(x+h) - f(x)}{F(x+h) - F(x)} = \frac{f'(x+\theta h)}{F'(x+\theta h)}$$

for some value of θ such that $0 < \theta < 1$. It is assumed that $F(x+h) \neq F(x)$.

To prove the theorem, let

$$\phi(\xi) = f(\xi) - f(x) - \frac{f(x+h) - f(x)}{F(x+h) - F(x)}\{F(\xi) - F(x)\};$$

and let it be assumed that $F(x+h) - F(x)$ is not zero. Since

$$\phi(x) = \phi(x+h),$$

and $\phi(\xi)$ satisfies the conditions of the mean value theorem, $\phi'(\xi)$ must vanish for some value $x + \theta h$, of ξ, interior to the interval $(x, x+h)$; $\phi'(\xi)$ exists, since $f'(\xi)$, $F'(\xi)$ cannot both be infinite. We have then

$$f'(x+\theta h) - \frac{f(x+h) - f(x)}{F(x+h) - F(x)} F'(x+\theta h) = 0,$$

from which the theorem follows, since $F'(x+\theta h)$ and $f'(x+\theta h)$ cannot be both infinite, or both zero. In the case in which $f(x+h) = f(x)$, we have

$$f'(x+\theta h) = 0,$$

for some suitable value of θ; and then, since $F(x+h) - F(x)$, $F'(x+\theta h)$ are not zero, the theorem still holds.

264. The last theorem may be applied to obtain a proof of the legitimacy, under certain conditions, of a well-known method of evaluating limits which appear in the so-called indeterminate forms $\frac{0}{0}$, $\frac{\infty}{\infty}$.

Let the two functions $f(x)$, $F(x)$ be both continuous at all points interior to the interval $(a, a+\beta)$, and let the limits $f(a+0)$, $F(a+0)$ be both unique, and have the value zero; if finite differential coefficients $f'(x)$, $F'(x)$ exist at every interior point of $(a, a+\beta)$, and if h $(> 0$, and $< \beta)$ can be so determined that, within $(a, a+h)$, $f'(x)$ and $F'(x)$ are not both zero, and not both infinite, together at any point, then if the limit

$$\lim_{h\sim 0} \frac{f'(a+h)}{F'(a+h)}$$

exist as a definite number, or be infinite with a fixed sign, the limit

$$\lim_{h\sim 0} \frac{f(a+h)}{F(a+h)}$$

also exists, and the two have the same value.

Moreover, if the first limit has no unique value, every value of the second limit is in the interval bounded by the upper and lower values of the first limit, and is a value of the first limit. Thus

$$\overline{\lim_{h\sim 0}} \frac{f'(a+h)}{F'(a+h)} \geqq \overline{\lim_{h\sim 0}} \frac{f(a+h)}{F(a+h)} \geqq \underline{\lim_{h\sim 0}} \frac{f(a+h)}{F(a+h)} \geqq \underline{\lim_{h\sim 0}} \frac{f'(a+h)}{F'(a+h)}.$$

The two functional values $f(a)$, $F(a)$ may both be defined to be zero, and thus the functions $f(x)$, $F(x)$ are continuous in any interval $(a, a+h)$, where $h < \beta$. We have then, from the extension of the mean value theorem (§ 263)

$$\frac{f(\alpha + h)}{F(\alpha + h)} = \frac{f'(\alpha + \theta h)}{F'(\alpha + \theta h)},$$

where θ is some number such that $0 < \theta < 1$, whose value depends in general on that of h. If there be assigned to h the values in a diminishing sequence which converges to zero, the corresponding sequence of values of θh also converges to zero.

In case $\lim\limits_{h\sim 0} \frac{f'(\alpha + h)}{F'(\alpha + h)}$ has a unique value, either finite, or infinite with fixed sign, it follows that $\lim\limits_{h\sim 0} \frac{f(\alpha + h)}{F(\alpha + h)}$ has the same value.

In case $\lim\limits_{h\sim 0} \frac{f'(\alpha + h)}{F'(\alpha + h)}$ has not a unique value, let the diminishing sequence $\{h_n\}$, converging to zero, be so chosen that $\frac{f(\alpha + h)}{F(\alpha + h)}$ has a single limit, as h goes through the sequence $\{h_n\}$, then $\lim\limits_{n\sim\infty} \frac{f'(\alpha + \theta_n h_n)}{F'(\alpha + \theta_n h_n)}$ has the same single limit, and this must be one of the values of $\lim\limits_{h\sim 0} \frac{f'(\alpha + h)}{F'(\alpha + h)}$. Therefore any limit of $\frac{f(\alpha + h)}{F(\alpha + h)}$, as $h \sim 0$, must also be a limit of $\frac{f'(\alpha + h)}{F'(\alpha + h)}$, as $h \sim 0$, and must lie in the interval which contains all the limits of $\frac{f'(\alpha + h)}{F'(\alpha + h)}$.

It may happen, in particular cases, that $\lim\limits_{h\sim 0} \frac{f(\alpha + h)}{F(\alpha + h)}$ has a unique value, whilst $\lim\limits_{h\sim 0} \frac{f'(\alpha + h)}{F'(\alpha + h)}$ is not unique.

For example, let $\alpha = 0$, $f(x) = x^2 \sin \frac{1}{x}$, $F(x) = e^x - 1$; then $\lim\limits_{x\sim 0} \frac{f(x)}{F(x)}$ has the unique value zero, whilst $\frac{f'(x)}{F'(x)}$, or $\left(2x \sin \frac{1}{x} - \cos \frac{1}{x}\right) e^{-x}$, has no unique limit, but its limits are in the interval $(-1, 1)$.

If $f(x)$, $F(x)$ have both differential coefficients $f'(x)$, $F'(x)$ at all points of the open interval $(\alpha, \alpha + \beta)$, and if $f(\alpha + 0)$, $F(\alpha + 0)$, both exist, and are infinite with a fixed sign; then, provided h $(> 0, < \beta)$ can be so determined that $f'(x)$, $F'(x)$ are not both zero, or both infinity, at any point in the open*

---

* This generalization of the conditions to be satisfied by the two functions in these theorems is due to W. H. Young; see *Proc. Lond. Math. Soc.* (2), vol. VIII (1909), p. 51. For a history of these theorems, see Pringsheim, *Encycl. d. math. Wissensch.* II A. I, p. 26. A detailed investigation of the theorems is given by Stolz, *Grundzuge*, vol. I, p. 77 *et seq.*; Stolz assumes that $F(x)$ is monotone in the interval.

interval $(a, a+h)$, *if* $\lim\limits_{h\sim 0}\dfrac{f'(a+h)}{F'(a+h)}$ *exists, as a finite number, or is infinite with a fixed sign, also* $\lim\limits_{h\sim 0}\dfrac{f(a+h)}{F(a+h)}$ *exists, and the two have the same value. Moreover, if the first limit has no unique value, every value of the second limit is in the interval bounded by the upper and lower values of the first limit, and is one of the values of the first limit.*

Consider the interval $(a+\delta_1, a+\delta_2)$ interior to $(a, a+h)$; we have then

$$\frac{f(a+\delta_2)-f(a+\delta_1)}{F(a+\delta_2)-F(a+\delta_1)}=\frac{f'(a+\delta_3)}{F'(a+\delta_3)},$$

where δ_3 is a number which lies between δ_1 and δ_2: it being assumed that $F(a+\delta_2) \neq F(a+\delta_1) \neq 0$. This equation may be written in the form

$$\frac{f(a+\delta_1)}{F(a+\delta_1)}=\frac{f(a+\delta_2)}{F(a+\delta_1)}+\frac{f'(a+\delta_3)}{F'(a+\delta_3)}\left\{1-\frac{F(a+\delta_2)}{F(a+\delta_1)}\right\}.$$

Taking δ_2 as fixed, if ϵ be an arbitrarily small positive number, a positive number δ' $(<\delta_2)$ can be so determined that for every value of δ_1 (>0) which is $<\delta'$, the inequalities

$$|F(a+\delta_1)|>\frac{1}{\epsilon}\,|f(a+\delta_2)|,\quad |F(a+\delta_1)|>\frac{1}{\epsilon}\,|F(a+\delta_2)|$$

are both satisfied; this follows from the fact that $F(a+0)$ is infinite with fixed sign.

We have now $\dfrac{f(a+\delta_1)}{F(a+\delta_1)}=\eta+(1+\zeta)\dfrac{f'(a+\delta_3)}{F'(a+\delta_3)}$,

where $|\eta|<\epsilon$, and $|\zeta|<\epsilon$, for all values of δ_1 such that $0<\delta_1<\delta'$; where δ_2 is fixed, and δ' $(<\delta_2)$ can then be determined.

If we assign to δ_2 the values in a sequence that converges to zero, δ' will have the values in a similar sequence.

If U, L denote the upper and the lower limit of $\dfrac{f'(a+h)}{F'(a+h)}$, as $h\sim 0$, we see that $\dfrac{f'(a+\delta_3)}{F'(a+\delta_3)}$ lies between $U+\epsilon$ and $L-\epsilon$, provided that δ_2 is sufficiently small.

Then $\dfrac{f(a+\delta_1)}{F(a+\delta_1)}$ is between $\epsilon+(1+\epsilon)(U+\epsilon)$ and $-\epsilon-(1-\epsilon)(L-\epsilon)$; and therefore, since ϵ is arbitrary, any value of $\lim\limits_{h\sim 0}\dfrac{f(a+h)}{F(a+h)}$ is in the interval (L, U).

If $U=L$, then $\lim\limits_{h\sim 0}\dfrac{f(a+h)}{F(a+h)}$ has the unique value $\lim\limits_{h\sim 0}\dfrac{f'(a+h)}{F'(a+h)}$.

Again, let the values of δ_1 be restricted to belong to a diminishing sequence $\{\delta_1^{(n)}\}$ converging to zero, and so chosen that $\dfrac{f(a+\delta_1)}{F(a+\delta_1)}$ converges to the unique limit A, as δ_1 goes through the values in the sequence $\{\delta_1^{(n)}\}$. Then, for all sufficiently large values of n, $\frac{f(\alpha+\delta_1^{(n)})}{F(\alpha+\delta_1^{(n)})} = A + \lambda$, where $|\lambda| < \epsilon$; and then $\frac{f'(\alpha+\delta_3)}{F'(\alpha+\delta_3)}$ has the values $\frac{A+\lambda-\eta}{1-\zeta}$, for all the values of δ_3 in a certain sequence. It follows that the upper and lower limits of $\frac{f'(\alpha+\delta_3)}{F'(\alpha+\delta_3)}$ for a certain sequence of values of δ_3, are in an interval

$$\left(\frac{A-2\epsilon}{1+\epsilon}, \frac{A+2\epsilon}{1-\epsilon}\right).$$

Since ϵ is arbitrary, it follows that $\frac{f'(\alpha+\delta_3)}{F'(\alpha+\delta_3)}$ converges to A.

265. Let $f(x)$, $F(x)$ be both zero at the point $x = x_1$, and let us suppose that $f'(x_1)$, $F'(x_1)$ both exist, and are finite, the latter not being zero. It is unnecessary to assume the existence of $f'(x)$, $F'(x)$ for $x \neq x_1$.

From the definition of $f'(x_1)$, $F'(x_1)$, we have

$$f(x_1+h) = h\{f'(x_1) + \rho(h)\},$$
$$F(x_1+h) = h\{F'(x_1) + \rho'(h)\},$$

where $\rho(h)$, $\rho'(h)$ converge to zero, as $h \sim 0$. Thence we have

$$\frac{f(x_1+h)}{F(x_1+h)} = \frac{f'(x_1)+\rho(h)}{F'(x_1)+\rho'(h)},$$

from which it follows that

$$\lim_{x\sim x_1} \frac{f(x)}{F(x)} = \frac{f'(x_1)}{F'(x_1)}.$$

This includes the case in which $f'(x_1) = 0$, $F'(x_1) \neq 0$, when the limit is zero.

Next, let it be assumed that $f(x_1)$, $f'(x_1)$, ... $f^{(n-1)}(x_1)$ (see § 270) all exist, and have the value 0; and that $F(x_1)$, $F'(x_1)$, ... $F^{(n'-1)}(x_1)$ all exist, and have the value 0. Also let it be assumed that $f^{(n)}(x_1)$, $F^{(n')}(x_1)$ exist, and are finite, neither of them being zero. It will also be assumed that, in some neighbourhood of the point x_1, the first $n-1$ differential coefficients of $f(x)$, and the first $n'-1$ differential coefficients of $F(x)$, all exist.

We have then,

$$\lim_{h\sim 0} \frac{f^{(n-1)}(x_1+h)}{h} = f^{(n)}(x_1); \text{ since } f^{(n-1)}(x_1) = 0.$$

It then follows, by the employment of the theorem of § 263, that

$$\lim_{h\sim 0} \frac{f(x_1+h)}{h^n} = \lim_{h\sim 0} \frac{f'(x_1+h)}{nh^{n-1}} = \lim_{h\sim 0} \frac{f''(x_1+h)}{n(n-1)h^{n-2}} = \ldots$$
$$= \lim_{h\sim 0} \frac{f^{(n-1)}(x_1+h)}{n!\,h} = \frac{f^{(n)}(x_1)}{n!}.$$

Similarly, we have $\lim_{h\sim 0}\frac{F(x_1+h)}{h^{n'}}=\frac{F^{(n')}(x_1)}{n'!}$; and thus we see that, since

$$\frac{f(x_1+h)}{F(x_1+h)}=h^{n-n'}\frac{f(x_1+h)/h^n}{F(x_1+h)/h^{n'}},$$

$$\lim_{h\sim 0}\frac{f(x_1+h)}{F(x_1+h)}=0,\text{ if } n'<n$$

$$=\pm\infty,\text{ if } n'>n$$

$$=f^{(n)}(x_1)/F^{(n)}(x_1),\text{ if } n'=n.$$

The following theorem has been established:

If $f(x)$, $F(x)$ are zero when $x=x_1$, and have, in a neighbourhood of that point, differential coefficients of the first $n-1$, and $n'-1$ orders respectively, which are all zero at the point $x=x_1$; and if, moreover, the differential coefficients of $f(x)$, $F(x)$, of orders n, n', respectively, exist and are finite, at the point $x=x_1$, both of them being different from zero, then $\lim_{x\sim x_1}\frac{f(x)}{F(x)}$ has a unique value 0, $\frac{f^n(x_1)}{F^n(x_1)}$, $\pm\infty$; according as $n'<n$, $n'=n$, or $n'>n$; the sign of $\pm\infty$ being that of $f^n(x_1)/F^{n'}(x_1)$.

If, for every value of x_1 in an open interval (α,β), $f'(x)$, $F'(x)$ are both continuous, and $F'(x)\neq 0$, we see, as in § 262, that

$$f(x+h)-f(x)=h\{f'(x)+\rho(x,h)\},$$
$$F(x+h)-F(x)=h\{F'(x)-\rho'(x,h)\},$$

where $\rho(x,h)$, $\rho'(x,h)$ converge to 0, as $h\sim 0$, uniformly for all points x in a closed interval (a,b) contained in (α,β). We then see that

$$\frac{f(x+h)-f(x)}{F(x+h)-F(x)}\text{ converges to }\frac{f'(x)}{F'(x)}$$

uniformly in the interval (a,b), provided $|F'(x)|$ exceed some fixed positive number, for all points in the interval.

266. *If the function $f(x)$ have a discontinuity of the second kind at the point a, at least on the side which is towards the interval $(a,a-h)$, but the function have a finite differential coefficient at every point of the interval $(a,a+h)$, except at the point a, then the differential coefficient has, in any arbitrarily small neighbourhood of a, every finite value.*

Let λ be any fixed number, positive, negative, or zero. The function $f(x)-\lambda x$ has a finite differential coefficient at all interior points of $(a,a+h)$, and it has a discontinuity of the second kind at the point a. If $\delta\,(<h)$ be arbitrarily small, in the interior of the neighbourhood $(a,a+\delta)$, $f(x)-\lambda x$ has maxima and minima, for it cannot be monotone in that interval; and at such points $f'(x)-\lambda=0$. Since λ is arbitrary, it thus appears that in the neighbourhood of a, $f'(x)$ has every finite value.

The mean value theorem $f(\alpha + h) - f(\alpha) = hf'(\alpha + \theta h)$, where $0 < \theta < 1$, affords information as to the existence and value of the derivative at α, on the right, provided $f(x)$ satisfies, in a neighbourhood of α on the right, the conditions under which the theorem holds. By considering both sides of α, information may be obtained as to the existence of a differential coefficient at α.

(1). If the function $f'(x)$ have a functional limit at α on the right, then

$$\frac{f(\alpha + h) - f(\alpha)}{h}$$

has a definite limit for $h \sim 0$, either finite, or infinite with fixed sign, and this is equal to that of $f'(x)$. It follows that, in this case, a derivative at α on the right exists, and is either finite, or infinite with fixed sign.

(2). If the function $f'(x)$ have no limit at α on the right, it may still happen that $f'(\alpha + \theta h)$ has a definite limit at α on the right, because $\alpha + \theta h$ is not necessarily capable of having all values within a neighbourhood of α. In this case, either (i), the derivative at α on the right may be definite, and lie between the upper and lower limits of $f'(x)$ at α on the right, or it may be equal to one or other of those limits; or (ii), there may be no definite derivative at α on the right, but $D^+f(\alpha)$, $D_+f(\alpha)$ may have different values, and these are certainly both finite in case the upper and lower functional limits of $f'(x)$ at α are both finite.

(3). The derivative on the right at α can only exist, and be infinite, (i), if $f'(x)$ have an infinite limit on the right at α, or (ii), if it have an infinite upper limit on the right at α. In either of the cases, (i) and (ii), $f'(x)$ may be everywhere finite within a neighbourhood of α on the right, or it may be infinite at some points in such a neighbourhood.

(4). If the derivative at α on the right exist, and be finite, then either (i), $f'(x)$ has a definite limit at α on the right, equal to the derivative at α, or (ii), $f'(x)$ has no definite limit at α on the right, but a sequence of points can be determined, of which α is the limiting point, such that the values of $f'(x)$ for points of that sequence converge to the value of the derivative at α. At points which do not belong to the sequence, the values of $f'(x)$ may be either finite or infinite.

267. *If $f(x)$ be continuous in a given closed interval, and have at every point, with the possible exception of an enumerable set G, a differential coefficient of value zero, the function is constant throughout the whole interval.*

At the points of G we may suppose it to be unknown whether a differential coefficient exists, or, if one does exist, what values it has.

A more general form of this theorem is obtained by considering not the differential coefficient, but any one of the four derivatives, thus:

If $f(x)$ be continuous in a given closed interval, and one of the four derivatives $D^+f(x)$, $D_+f(x)$, $D^-f(x)$, $D_-f(x)$, be such that it is zero at every point of the interval, with the exception of points belonging to an enumerable set G, at which nothing is known as to its value, then the function is constant throughout the interval.

To prove the generalized theorem for the case of the function $D^-f(x)$, suppose that, if possible, $f(x) - f(a)$ has at some point x_1 a value different from zero, say the positive value p; and let $\phi(x, k)$ denote

$$f(x) - f(a) - k(x - a).$$

Then $\phi(a, k) = 0$, $\phi(x_1, k) = p - k(x_1 - a)$. Choose any fixed positive number $q < p$, then $\phi(x_1, k) > q$, provided $k < \frac{p - q}{x_1 - a}$, or say $k < K$. Since $\phi(x, k)$ is continuous in (a, b), and $\phi(a, k)$ is zero, whilst $\phi(x_1, k) > q$, there exists an upper limit of those values of x between a and x_1, for which $\phi(x, k) \leqq q$, and this upper limit is attained for some value ξ, of x, which is such that $\xi < x_1$, and $\phi(\xi, k) = q$. Since $\phi(\xi - h, k) > q$, provided $0 < h \leqq x_1 - \xi$, we see that, since $\frac{\phi(\xi - h, k) - \phi(\xi, k)}{h}$ is positive, $D^+\phi(\xi, k)$ is positive, if it be not zero. Now if ξ were a point not belonging to G, the value of $D^+\phi(\xi, h)$ would reduce to $-k$; and therefore ξ must belong to G.

The number q being fixed, ξ depends only on k; and, corresponding to a given value of ξ, there is only one value of k; for

$$\phi(\xi, k) - \phi(\xi, k') = (k' - k)(\xi - a),$$

which cannot vanish unless $k = k'$, since $\phi(a, k)$ is zero, and therefore $< q$. For a given value of k, the corresponding number of values of ξ, all of which necessarily belong to G, must be either finite, or enumerably infinite, since every part of an enumerable aggregate is either finite or enumerable. Therefore, to each value of k, in the continuous interval $(\alpha, K - \beta)$, there corresponds a finite, or enumerable, set of values of ξ, and it would hence follow that the continuum $(\alpha, K - \beta)$ is itself enumerable, which we know is not the case. It has thus been shewn that, for no point can $f(x) - f(a)$ have a positive value; and similarly, by considering

$$f(x) - f(a) + k(x - a),$$

it can be shewn that $f(x) - f(a)$ can nowhere have a negative value: hence $f(x) = f(a)$ throughout the whole interval (a, b). The case in which one of the other three derivatives vanishes, except at points of G, can be treated in a similar manner.

A continuous function can exist, which is constant in each interval contiguous to a non-dense closed set G, and is not everywhere constant, provided G is unenumerable; but if G be enumerable, the function must everywhere have the same constant value.

The following theorem* which is of importance in the theory of Integration will now be established:

If two functions be each continuous in a given closed interval, and if, of one of the four derivatives it be known that, for the two functions, this derivative has equal finite values at each point of the interval, with the exception of an enumerable set of points at which nothing is known as regards the two derivatives, then the two functions differ from one another only by a constant, which must be the same for the whole interval.

It must first be observed that the proof of the preceding theorem suffices to shew that, if $D^+f(x) \leqq 0$, at every point of (a, b) not belonging to the set G, then $f(x) - f(a) \leqq 0$, for every point x of the interval. Similarly, if $D^+f(x) \geqq 0$, everywhere in the interval, except at the points of G, then $f(x) - f(a) \geqq 0$, at every point of the interval.

If now $f_1(x), f_2(x)$ be two continuous functions such that

$$D^+f_1(x) = D^+f_2(x)$$

at every point of (a, b) not belonging to G, let $f(x) = f(x_1) - f(x_2)$. If ϵ be an arbitrarily small positive number, then for any point x not belonging to G, the condition

$$\frac{f_1(x+h) - f_1(x)}{h} > D^+f_1(x) - \epsilon$$

is satisfied for a set of positive values of h which are arbitrarily small. Also we have, for all sufficiently small values of h,

$$\frac{f_2(x+h) - f_2(x)}{h} < D^+f_2(x) + \epsilon;$$

hence, since $D^+f_1(x) = D^+f_2(x)$, we see that $\frac{f(x+h) - f(x)}{h} > -2\epsilon$, for all values of h belonging to some set. It follows that $D^+f(x) > -2\epsilon$, and thence† that $D^+f(x) \geqq 0$, since ϵ is arbitrary. By interchanging $f_1(x)$ and $f_2(x)$, we see that $D^+\{-f(x)\} \geqq 0$. From these two results we deduce that $f(x) - f(a) \geqq 0$, and that $f(a) - f(x) \geqq 0$, throughout the interval (a, b); therefore $f(x)$ is everywhere equal to $f(a)$, and thus the theorem is established.

Examples have been given by Hahn‡ and by Ruziewicz§ of continuous functions which have everywhere in an interval the same differential coefficient (not everywhere finite) but the difference of which is not constant in the interval.

---

* Scheeffer, *Acta Math* vol. v (1884), p. 283.

† It is erroneously stated by Dini, that $D^+f(x) = 0$. See *Grundlagen*, p. 275.

‡ *Monatshefte f. Math. u. Physik*, vol. xvi (1905), p. 16.

§ *Fundamenta Mat.* vol. i (1920), p. 148.

268. At a point x, at which the continuous function $f(x)$ is a maximum, since, for a sufficiently small neighbourhood of such a point x, the differences

$$f(x+h)-f(x),\ f(x-h)-f(x)$$

are both negative or zero, for all points $x \pm h$ in the neighbourhood, it is clear that each of the derivatives $D^+f(x)$, $D_+f(x)$ is either negative or zero, and that each of the derivatives $D^-f(x)$, $D_-f(x)$ is either positive or zero. In case the function possess definite derivatives on the right and on the left at the point x, the first of these is zero or negative, or possibly $-\infty$, while the second is zero or positive, or possibly $+\infty$.

If, at the point x, a definite differential coefficient exist, it must consequently be zero. In the case of a minimum the corresponding statements hold, where the positive sign takes the place of the negative one, and the reverse. The following theorem has now been established:

If a continuous function possess a differential coefficient at a point x at which the function is a maximum or minimum, then the differential coefficient at x must be zero.

If a function $f(x)$ have, at x_1, a differential coefficient not equal to zero, there is a neighbourhood of x_1 in which $f(x)$ is monotone. This may be stated in the form that, corresponding to a suitable value of h, a number k can be determined, so that, to each value of y in the interval

$$(y_1-k,\ y_1+k),$$

where $y_1=f(x_1)$, there corresponds one, and only one, value of x in the interval $(x_1-h,\ x_1+h)$.

FUNCTIONS WITH LINES OF INVARIABILITY

269. A continuous function may be such that, in the interval (a, b), there exists an everywhere dense set of non-overlapping intervals, each one of which is a line of invariability of the function. Within each interval of the set, the function has its differential coefficient equal to zero: it therefore follows from the theorem in § 267, that the closed set of points, of which the given set of intervals is the complementary set, cannot be an enumerable set, otherwise the function would be constant in the whole interval (a, b). It is further clear that no two of the intervals can abut on one another; for the condition of continuity of the function at their common end-point would ensure that the values of the function in the two intervals were the same, and thus the two intervals would really belong to the same line of invariability. It follows that the end-points and external points of an everywhere dense set of lines of invariability of a continuous function must form a perfect non-dense set of points.

That a continuous function with an everywhere dense set of lines of invariability can actually exist can be easily shewn as follows:—Make the points of a non-dense perfect set correspond in order to the points of a continuous interval (a, b), then, as has been shewn in § 163, the correspondence may be such that the whole of a complementary interval of the perfect set corresponds to one point of the continuous interval. If a continuous function be defined for the continuous interval, we may define a new function which has at each point of the perfect set the same value as the original function has at the corresponding point of the continuous interval; and since all the points of a complementary interval of the perfect set correspond to the same point of the continuous interval, the new function is such that it has an everywhere dense set of lines of invariability.

EXAMPLES

1. Take* the non-dense perfect set defined in Ex. 1, § 83, by

$$x = \frac{c_1}{3} + \frac{c_2}{3^2} + \dots + \frac{c_n}{3^n} + \dots,$$

where every c_n is either 0 or 2. A complementary interval has as its end-points

$$\frac{c_1}{3} + \frac{c_2}{3^2} + \dots + \frac{c_{n-1}}{3^{n-1}} + \frac{1}{3^n}, \quad \frac{c_1}{3} + \frac{c_2}{3^2} + \dots + \frac{c_{n-1}}{3^{n-1}} + \frac{2}{3^n},$$

which may be denoted by (a_ν, b_ν). Let the function $f(x)$ be defined as follows.—For a point x of the interval $(0, 1)$ belonging to the perfect set, let

$$f(x) = \frac{1}{2}\left(\frac{c_1}{2} + \frac{c_2}{2^2} + \dots + \frac{c_n}{2^n} + \dots\right);$$

when x is in the interval (a_ν, b_ν), let $f(x) = f(a_\nu) = f(b_\nu)$. The function $f(x)$ so defined is continuous, and varies from 0 to 1, and is constant in each of the intervals (a_ν, b_ν) complementary to the non-dense perfect set.

2.† Let the numbers in the interval $(0, 1)$ be expressed in a scale $n \equiv 2m - 1$, of odd degree; thus $x = \frac{a_1}{n} + \frac{a_2}{n^2} + \dots$, where $0 \leqq a_r < n$, and the number of digits a_r is finite or infinite. For any number x represented in this manner, for which all the a's are even integers, let $f(x)$ equal $\frac{1}{2}\left(\frac{a_1}{m} + \frac{a_2}{m^2} + \dots\right)$. In case any of the a's are odd, let a_k be the first one which is odd, and let $f(x)$ then equal $\frac{1}{2}\left(\frac{a_1}{m} + \frac{a_2}{m^2} + \dots + \frac{a_{k-1}}{m^{k-1}}\right) + \frac{1}{2}\left(\frac{a_k + 1}{m^k}\right)$. This function $f(x)$ is continuous, and varies from 0 to 1; for an infinite set of points it has no differential coefficient, and for all other values of x, $f'(x) = 0$.

THE SUCCESSIVE DIFFERENTIAL COEFFICIENTS OF A CONTINUOUS FUNCTION

270. If a continuous function $f(x)$, defined for the closed interval (a, b), have at every point a differential coefficient $f'(x)$, which is itself continuous throughout the interval, the function $f'(x)$ may itself have a differential coefficient $f''(x)$, which is called the second differential coefficient, or the second derivative, of $f(x)$.

---

* Cantor, *Acta Math.* vol. IV (1884), p. 386. See also Scheeffer, *Acta Math.* vol V (1884), p. 289.

† Gravé, *Comptes Rendus*, vol. CXXVII (1898), p. 1005.

The second differential coefficient $f''(x_1)$, of $f(x)$, at a point x_1, at which $f'(x_1)$ exists as a definite number, may be defined as

$$\lim_{h\sim 0}\frac{f'(x_1+h)-f'(x_1)}{h},$$

when this limit exists, as a finite number, or is infinite with a fixed sign. It is not necessary for the existence of the second differential coefficient $f''(x_1)$, that $f'(x_1+h)$ should have a definite value, for all points x_1+h in a neighbourhood of x_1.

It may happen that the four ratios

$$\frac{1}{h}\{D^+f(x_1+h)-f'(x_1)\},\quad \frac{1}{h}\{D_+f(x_1-h)-f'(x_1)\},$$

$$\frac{1}{h}\{D^-f(x_1+h)-f'(x_1)\},\quad \frac{1}{h}\{D_-f(x_1-h)-f'(x_1)\},$$

all have one and the same limit, finite, or infinite, when $h\sim 0$; and that this should be the case is sufficient for the existence of $f''(x_1)$, as here defined.

The second differential coefficient $f''(x_1)$ is thus defined as the repeated limit

$$\lim_{h\sim 0}\lim_{k\sim 0}\frac{f(x_1+h+k)-f(x_1+h)-f(x_1+k)+f(x_1)}{hk};$$

when this repeated limit exists; it being assumed that

$$\lim_{k\sim 0}\frac{f(x_1+k)-f(x_1)}{k}$$

has a definite value.

A more restricted definition of $f''(x_1)$ is frequently employed, in which it is assumed that $f'(x)$ exists as a definite number, not only at x_1, but in a neighbourhood of x_1. In accordance with this more restricted definition, it is not sufficient for the existence of $f''(x_1)$ that the repeated limit

$$\lim_{h\sim 0}\lim_{k\sim 0}\frac{f(x_1+h+k)-f(x_1+h)-f(x_1+k)+f(x_1)}{hk}$$

should have a definite, finite value, or be infinite with fixed sign, even if it be assumed that $f'(x_1)$ exists.

For the sake of generality the less restricted definition of $f''(x_1)$, given above, will be here employed.

It is easily seen how the definition may be extended to the case of a differential coefficient $f^{(n)}(x)$ of any order n.

271. In case the function $f(x)$ possesses a finite differential coefficient, of order n, at the point $x=x_1$, and in which there is a neighbourhood of that point in which $f'(x), f''(x), \dots, f^{(n-1)}(x)$ all exist, let

$$f(x_1+h)-f(x_1)-hf'(x_1)-\frac{h^2}{2!}f''(x_1)-\dots-\frac{h^{n-1}}{(n-1)!}f^{(n-1)}(x_1)$$

be denoted by $\phi(h)$. Since $\phi(h), \phi'(h), \ldots, \phi^{(n-1)}(h)$ all exist, and are zero, when $h = 0$, it follows from the theorem of § 263 that

$$\lim_{h\sim 0} \frac{\phi(h)}{h^n} = \frac{1}{n!}\phi^{(n)}(0) = \frac{1}{n!}f^{(n)}(x_1);$$

and then that $\phi(h) = \frac{h^n}{n!}\{f^{(n)}(x_1) + \rho(h)\}$, where $\rho(h)$ converges to 0, as $h \sim 0$. We have thus the following theorem:

If $f(x)$ possesses an n^{th} differential coefficient at $x = x_1$, and if the first $n-1$ differential coefficients exist everywhere in some neighbourhood of the point $x = x_1$, then

$$f(x_1+h) = f(x_1) + hf'(x_1) + \frac{h^2}{2!}f''(x_1) + \ldots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(x_1) + \frac{h^n}{n!}\{f^{(n)}(x_1) + \rho(h)\},$$

where $\lim\limits_{h\sim 0} \rho(h) = 0$.

We find, by employing this theorem, that when $f''(x_1)$ exists, and $f'(x)$ exists in a neighbourhood of x_1,

$$\frac{f(x_1+h+k) - f(x_1+h) - f(x_1+k) + f(x_1)}{hk} = f''(x_1) + \sigma(h, k),$$

where the double limit $\lim\limits_{\substack{h\sim 0\\ k\sim 0}} \sigma(h, k)$ exists and has the value zero.

It follows that the double limit of the expression

$$\frac{f(x_1+h+k) - f(x_1+h) - f(x_1+k) + f(x_1)}{hk}$$

exists and is equal to $f''(x_1)$. The converse does not in general hold.

In particular the limits

$$\lim_{h\sim 0}\frac{f(x_1+2h) - 2f(x_1+h) + f(x_1)}{h^2}, \quad \lim_{h\sim 0}\frac{f(x_1+h) + f(x_1-h) - 2f(x_1)}{h^2}$$

exist and have the value $f''(x_1)$.

The converse of this does not hold; for either of these limits may exist, and yet $f''(x_1)$ need not exist, nor even $f'(x_1)$. An illustration of this is the case of the function defined by $f(0) = 0, f(x) = x\sin^2\frac{1}{x}$, for $x^2 > 0$; at the point $x = 0$, $f'(0)$ does not exist, and yet $\lim\limits_{h\sim 0}\frac{f(h) - 2f(0) + f(-h)}{h^2} = 0$.

272. The following theorem, due to Schwarz*, is of fundamental importance in the theory of Fourier's series.

If, in an interval (α, β), in which $f(x)$ is continuous,

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

---

* *Gesamm. Abhandlungen*, vol. II (1890), p. 341.

converge, for each value of x in (α, β), to the limit zero, for $h \sim 0$, then the function $f(x)$ is a linear function in the whole interval, and consequently $f'(x), f''(x)$ everywhere exist, and the latter is everywhere zero.

Let us consider the function

$$\phi(x) = \pm \left\{ f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}[f(\beta) - f(\alpha)] \right\} - k^2 (x-\alpha)(x-\beta),$$

where k is a constant. The function $\phi(x)$, whichever sign be taken, is continuous in (α, β), and vanishes at α and β. We find at once

$$\lim_{h=0} \frac{\phi(x+h) - 2\phi(x) + \phi(x-h)}{h^2} = 2k^2;$$

and therefore, for each value of x in (α, β), a positive number ϵ can be found, such that $\phi(x+h) - 2\phi(x) + \phi(x-h)$ is positive and greater than zero, for all values of h which are numerically less than ϵ.

If $\phi(x)$ can be anywhere positive in (α, β), there must be a point x_1 at which it has the greatest positive value, and this point is neither α nor β, since $\phi(\alpha)$, $\phi(\beta)$ both vanish. If η be sufficiently small,

$$\phi(x_1+\eta) - \phi(x_1) \leq 0, \quad \phi(x_1-\eta) - \phi(x_1) \leq 0,$$

hence
$$\phi(x_1+\eta) - 2\phi(x_1) + \phi(x_1-\eta)$$

is, for all sufficiently small values of η, either negative or zero, which is contrary to what was shewn above. It follows that $\phi(x)$ is everywhere negative in (α, β), and cannot be zero except at α and β.

This holds whichever sign be taken in defining $\phi(x)$. Now

$$k^2(x-\alpha)(x-\beta)$$

is always negative, except at α and β, and may be taken to have its numerically greatest value as small as we please, since k is at our choice. It follows that

$$f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}[f(\beta) - f(\alpha)]$$

can nowhere in the interval be different from zero: for, if at any point it had a value p, by choosing k such that $k^2(x-\alpha)(x-\beta)$ is numerically everywhere $< |p|$, the function $\phi(x)$ could be made positive at the point, by proper choice of the ambiguous sign. It has thus been shewn that $f(x)$ is linear in (α, β).

273. Schwarz's theorem can be extended to the case in which there is an enumerable set of points in the interval (α, β), at which it is not known that the limit in question exists, or is zero, provided a certain condition be satisfied at each point of the enumerable set. The following theorem will be established:

If, in an interval (α, β), in which $f(x)$ is continuous, the expression

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

converge, for each value of x in (α, β), to the limit zero, for $h \sim 0$, except that, for an enumerable set of points G, this is not known to be the case, then, provided that at each point x, of G, the expression

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h}$$

converge to the limit zero, for $h \sim 0$, the function $f(x)$ is a linear function in the whole interval (α, β).

It should be observed that the condition

$$\lim_{h=0} \frac{f(x+h) - 2f(x) + f(x-h)}{h} = 0$$

is certainly satisfied at any point x at which the differential coefficient $f'(x)$ exists, and is finite. Moreover, whenever the condition is satisfied, $f(x)$ has its derivatives symmetrical as regards the right and left of the point x.

To prove the theorem, let it be assumed that

$$f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}\{f(\beta) - f(\alpha)\}$$

has a positive value p, at some point x_1 interior to (α, β); and let

$$\phi(x, k) = f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}\{f(\beta) - f(\alpha)\} + k(x-\alpha)^2,$$

where k is a positive number. We have

$$\phi(\alpha, k) = 0,\ \phi(\beta, k) = k(\beta-\alpha)^2, \text{ and } \phi(x_1, k) = p + k(x_1-\alpha)^2;$$

and hence, provided $k < \dfrac{p}{(\beta-\alpha)^2 - (x_1-\alpha)^2} = K$,

the number $\phi(x_1, k)$ is greater than $\phi(\beta, k)$, and than $\phi(\alpha, k)$. We shall suppose k to be so chosen that this condition is satisfied; it then follows that $\phi(x, k)$ has a maximum between α and β. The absolute maximum value of $\phi(x, k)$ may be attained once, or a finite number of times, or an infinite number of times, in the interval (α, β).

The points x at which this maximum is attained have an upper extreme $\bar{x}$ ($< \beta$), which must itself be a point at which the maximum of $\phi(x, k)$ is attained, as is seen, in the case in which $\bar{x}$ is an upper limit, from the condition of continuity of the function. We have therefore

$$\phi(\bar{x}+h, k) - \phi(\bar{x}, k) \leq 0, \text{ and } \phi(\bar{x}-h, k) - \phi(\bar{x}, k) \leq 0,$$

if h be sufficiently small; from which we conclude that, in case

$$\lim_{h \sim 0} \frac{\phi(\bar{x}+h, k) - 2\phi(\bar{x}, k) + \phi(\bar{x}-h, k)}{h^2} \text{ exist,}$$

its value is ≤ 0. It follows that $\bar{x}$ must belong to G; because the value of this limit is $2k$, and therefore > 0, for any point which does not belong to G. Since $\bar{x}$ is a point of G, we have

$$\lim_{h\sim 0}\left\{\frac{\phi(\bar{x}+h, k)-\phi(\bar{x}, k)}{h}+\frac{\phi(\bar{x}-h, k)-\phi(\bar{x}, k)}{h}\right\}=0;$$

and since the two fractions have the same sign, it follows that

$$\lim_{h\sim 0}\frac{\phi(\bar{x}+h, k)-\phi(\bar{x}, k)}{h}=0, \text{ and } \lim_{h\sim 0}\frac{\phi(\bar{x}-h, k)-\phi(\bar{x}, k)}{h}=0.$$

From this result we deduce that

$$\lim_{h\sim 0}\frac{f(\bar{x}+h)-f(\bar{x})}{h}=\lim_{h\sim 0}\frac{f(\bar{x}-h)-f(\bar{x})}{-h}=\frac{f(\beta)-f(\alpha)}{\beta-\alpha}-2k(\bar{x}-\alpha).$$

To each value of k in the interval $(0, K)$, there corresponds one value of $\bar{x}$, and it is impossible that the same value of $\bar{x}$ can correspond to two different values k_1, k_2 of k. For if this were the case, we should have

$$k_1(\bar{x}-\alpha)=k_2(\bar{x}-\alpha),$$

and therefore $k_1 = k_2$, since $\bar{x} > \alpha$. Now it is impossible that the set of points k, interior to the interval $(0, K)$, can be such that to each such point there corresponds a distinct point $\bar{x}$ belonging to the enumerable set G. We conclude that it is impossible that

$$f(x)-f(\alpha)-\frac{x-\alpha}{\beta-\alpha}\{f(\beta)-f(\alpha)\}$$

can have a positive value p at any point x_1 of the interval (α, β): and it can be shewn in a similar manner that there can be no negative value of the same function in the interval. It follows that the function must everywhere be zero, and therefore that $f(x)$ is linear in the interval (α, β).

274. Let us suppose that the continuous function $f(x)$ is linear in each interval contiguous to an enumerable closed set G, contained in (α, β), and that it possesses everywhere in the interval (α, β) a finite differential coefficient.

In this case both the conditions of the theorem of § 273 are satisfied, and therefore the function is linear in the whole interval (α, β); we have then the theorem:

If $f(x)$ be a continuous function possessing everywhere in the interval (α, β) a finite differential coefficient, and the function be linear in each one of an everywhere dense set of intervals complementary to an enumerable closed set of points G, then $f(x)$ is a linear function in (α, β).

If the closed set of points G were unenumerable, the preceding reasoning would no longer be applicable, except that, at an isolated point of G, it would establish that the linear functions in the two intervals which abut on one another at the isolated point must be identical. Confining therefore our attention to the case in which G is a perfect set, we see that a continuous function possessing everywhere a finite differential coefficient may exist, which is linear in each sub-interval complementary to a non-dense perfect set of points contained in the interval for which the function is defined, and yet the function need not be linear in the whole interval.

The existence of such functions will be effectively established in Chapter VII, where it will be shewn that they may be obtained by the integration of continuous functions which have an everywhere dense set of lines of invariability.

OSCILLATING CONTINUOUS FUNCTIONS

275. Let us suppose that the continuous function $f(x)$ has no lines of invariability in the interval (α, β), and that everywhere in this interval it has a finite differential coefficient. If, within (α, β), there be a maximum or minimum of $f(x)$, then at such a point $f'(x)$, which exists and is finite, must be zero. If the maxima and minima in (α, β) be everywhere dense, then $f'(x)$ vanishes at every point of the everywhere dense set; and if $f'(x)$ were continuous throughout (α, β) it would follow that it was everywhere zero, which would be contrary to the hypothesis that (α, β) is not a line of invariability.

It follows from this that, *if in an interval* (α, β), *which contains no lines of invariability of the continuous function* $f(x)$, *the differential coefficient* $f'(x)$ *everywhere exists, and is continuous, there must be in the interval an everywhere dense set of sub-intervals in each of which the function is monotone.*

We have further the following theorem:

If $f(x)$ *be continuous in* (α, β), *and have no lines of invariability, but have an everywhere dense set of maxima and minima, there must be in the interval an everywhere dense set of points at each of which* $f'(x)$ *either does not exist, or does exist and is discontinuous.*

A continuous function $f(x)$ which, in a given interval (α, β), has no lines of invariability, but has an everywhere dense set of maxima and minima, is said to be a continuous function which is *everywhere oscillating* in the interval (α, β). Such a function cannot have a differential coefficient which is continuous throughout the interval.

The continuous functions which are everywhere oscillating in an interval may be divided into two classes.

(1). The function may be such that, if the constants l, m be properly chosen, the function $f(x) + lx + m$ is monotone in the interval. In this case $f(x)$ is expressible as the difference of two monotone functions, and thus belongs to the class of functions with bounded variation. These functions may be said to be of the *first species*, or to be functions with *removable* oscillations.

(2). Such functions as do not belong to (1) may be said to be of the *second species*, or to be functions with *irremovable* oscillations.

In order to bring to light the essential distinction between the two classes of functions, as exhibited by the properties of their derivatives, we first of all remark that, if $D_-f(x)$ have a positive lower boundary c, for all points x in the interval (α, β), then at each point $f(x-h) - f(x)$ is essentially positive for all positive values of h which are less than some number δ, dependent on x; hence the function is monotone in the interval. The function would also be monotone in case the specified condition were that $D^+f(x)$ has a negative upper boundary for all values of x in (α, β).

Now suppose that $D_+f(x)$ has a definite negative lower boundary in (α, β); let this be $-c$, and consider the function $\phi(x) = f(x) - lx - m$, where $l > c$; we have then $D_+\phi(x) = l - D_-f(x) \geq l - c$; hence the function $\phi(x)$ is monotone in (α, β). Thus $f(x)$ is expressible as the difference of the two monotone functions $\phi(x)$ and $lx - m$. Similarly, if we had taken the condition that $D^-f(x)$ has a definite positive upper limit c, the function $f(x) + lx + m$, where $l < -c$ could be shewn to be monotone.

It is clear that, instead of the linear function $lx - m$, we might have used any continuous differentiable function whose differential coefficient was $> c$, or $< -c$, throughout the interval, in the two cases.

The argument would have been unaltered if it had been assumed that there were a finite or infinite set of lines of invariability in (α, β).

It has thus been shewn that:

If the continuous function $f(x)$ be such that either $D_-f(x)$ has a negative lower boundary for all values of x in (α, β), or that $D^-f(x)$ has a positive upper boundary, then all maxima and minima of the oscillating function $f(x)$ are removed by adding to $f(x)$ a properly chosen linear function, and thus the function is of the first species, and is of bounded variation.

In particular, the conditions of the theorem are satisfied if the derivative, on one side, without necessarily having a definite value at any point, be such that for the whole interval it is numerically less than some fixed positive number.

A function, such that for a given interval,

$$|D^+f(x)|, \quad |D_+f(x)|, \quad |D^-f(x)|, \quad |D_-f(x)|$$

are all less than some fixed number, is said to be a function *with bounded derivatives*. Such a function has bounded variation in the interval, and if it be everywhere oscillating, it is of the first species.

If all the four derivatives are, at every point, numerically less than the positive number M, in which case the function is necessarily continuous, a neighbourhood of any point x may be so determined that, if ξ be any point in that neighbourhood, $|f(\xi) - f(x)| \leq M|\xi - x|$. Hence the fluctuation in the neighbourhood δ is $\leq 2M\delta$.

By the Heine-Borel theorem all the points of (a, b) are interior to a finite number of the intervals δ. It follows that the interval (a, b) may be divided into a finite number of parts, each of length less than a prescribed number d, such that the sum of the fluctuations in those parts is $\leq 2M(b-a)$. Since d is arbitrarily small, it follows that the total fluctuation of $f(x)$ in (a, b) cannot exceed $2M(b-a)$.

It is not sufficient, in order that a function may be of limited total fluctuation, that, at each point, all four derivatives are finite; but it has been shewn to be sufficient that their values in the whole interval should be bounded.

276. Let us suppose that, for a set of points G, everywhere dense in (a, b), progressive and regressive derivatives at a point of G exist, and are infinite, but of opposite signs. At any point x_0, of G, a neighbourhood can be found, containing x_0, such that, for any point x in it, $f(x) - f(x_0)$ is of fixed sign for the whole neighbourhood, and is never zero, except when $x = x_0$; it follows that x_0 is a proper maximum, or minimum, of the function.

It will be shewn that, in any interval (α, β) contained in (a, b), there is an infinite number of points at which the function has the same value. Let ξ be a maximum point of $f(x)$, within (α, β), and let $(\xi - \eta, \xi + \epsilon)$ be the greatest interval enclosing ξ, within which $f(x) - f(\xi)$ is negative; suppose that the absolute minimum of the function for this interval is in $(\xi - \eta, \xi)$; taking a maximum point ξ_1 in the interval $(\xi, \xi + \epsilon)$, then in $(\xi - \eta, \xi)$ there is a point ξ_1' at which $f(\xi_1') = f(\xi_1)$, since $f(\xi_1)$ lies between the greatest and least values of the continuous function in $(\xi - \eta, \xi)$.

Now there is a maximum interval $(\xi_1 - \eta_1, \xi_1 + \epsilon_1)$ for the point ξ_1, and this lies within $(\xi, \xi + \epsilon)$; and in this interval we may as before find a maximum point ξ_2, such that a point ξ_2' also exists within the interval, for which $f(\xi_2) = f(\xi_2')$. There is also a point ξ_2'' in $(\xi - \eta, \xi)$, such that

$$f(\xi_2'') = f(\xi_2') = f(\xi_2).$$

We may proceed in this manner, until we find n points

$$\xi_{n-1}, \xi'_{n-1}, \dots \xi_{n-1}^{(n-1)},$$

such that
$$f(\xi_{n-1}) = f(\xi'_{n-1}) = \dots = f(\xi_{n-1}^{(n-1)}).$$

Now let ξ_ω be a limiting point of $\xi_1, \xi_2, \xi_3, \dots \xi_n, \dots$; and let ξ_ω' be a limiting point of $\xi_1', \xi_2', \dots$, and ξ_ω'' be a limiting point of $\xi_2'', \xi_3'', \dots$; then

$$f(\xi_\omega) = f(\xi_\omega') = f(\xi_\omega'') = \dots.$$

Thus the points $\xi_\omega, \xi_\omega', \xi_\omega'', \dots$ form an infinite set, in (α, β), at which the functional values are the same.

The points $\xi_\omega, \xi_\omega', \xi_\omega'', \dots$ have a limiting point ξ_0, at which the functional value is the same as for the set itself; therefore

$$\frac{f(\xi_0) - f(\xi_\omega)}{\xi_0 - \xi_\omega} = \frac{f(\xi_0) - f(\xi_\omega')}{\xi_0 - \xi_\omega'} = \dots = 0;$$

hence, at ξ_0, either the derivative is determinate, and equal to zero, or else it is indeterminate, with zero lying between its upper and lower limits. Thus it has been shewn that*:

If a continuous function have an everywhere dense set of points at which there are progressive and regressive derivatives that are infinite and of opposite signs, there is an everywhere dense set of points at each of which the derivative is either indeterminate or else zero. Thus a continuous function cannot at all points have infinite progressive and regressive derivatives of opposite signs.

If we apply the above theorem to the function $f(x) - cx$, where c is a prescribed constant, then, since $f(x) - cx$ has an infinite derivative at the same points as those for which $f(x)$ has an infinite derivative, we obtain the following theorem:

If the continuous function, $f(x)$, have at an everywhere dense set of points infinite progressive and regressive derivatives of opposite signs, there is an everywhere dense set of points at each of which the derivative either has the prescribed value c, or is indeterminate, and such that c lies between its upper and lower limits.

In geometrical language, a point x, at which there are infinite progressive and regressive derivatives, of opposite signs, is a point at which the curve $y = f(x)$ has a cusp, with its tangent perpendicular to the x-axis. At a maximum the cusp points in the positive direction of the y-axis, the progressive derivative being $-\infty$, and the regressive derivative $-\infty$. At a minimum the cusp points in the negative direction of the y-axis, the progressive derivative being $+\infty$, and the regressive derivative being $-\infty$.

PROPERTIES OF INCREMENTARY RATIOS

277. If x_1 be a point of the interval (a, b), in which $f(x)$ is defined, the function $\dfrac{f(x) - f(x_1)}{x - x_1}$, for points x such that $x_1 < x \leq b$, may be called the incrementary ratio at x_1 on the right; it may be denoted by $I(x, x_1)$; and in case $f(x)$ be a continuous function, this incrementary ratio is also continuous at every point of its domain. This incrementary function has an upper and a lower boundary for its whole domain $(x_1 < x \leq b)$; these upper and lower boundaries may be denoted by $U(x_1)$, $L(x_1)$, and either of them may be finite or infinite; however $U(x_1)$ can only be infinite with the positive sign, and $L(x_1)$ only with the negative sign. $U(x_1)$, $L(x_1)$ being regarded as functions of x_1, defined for every point of (a, b), except the point b, the function $U(x_1)$ has a finite or infinite upper boundary for its whole domain, which we may denote by U; and the function $L(x_1)$ has a finite or infinite lower boundary for its whole domain, which we may denote by L. There exist therefore two numbers U, L, which may have the improper values $+\infty$, $-\infty$ respectively, such that

---

* Konig, *Monatshefte f. Math. u. Physik*, vol. I (1890), p. 7. The above proof is that given by Schoenflies, *Bericht*, vol. I, p. 160.

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1},$$

for every pair of values of x_1, x_2, where $x_2 > x_1$, always lies between them, or is equal to one of them.

The incrementary ratio on the left of a point can be defined in a similar manner; and we thus define two functions $U'(x_1)$, $L'(x_1)$, at x_1, as the upper and lower boundaries of these incrementary ratios.

It is easily seen that U', the upper boundary of $U'(x)$ in the interval (a, b), is identical with U, and that L', the lower boundary of $L'(x)$, is identical with L. Thus U, L are the upper and lower boundaries of $I(x_1, x_2)$ for every possible pair of points (x_1, x_2) in the interval (a, b).

278. *When the function $f(x)$ is continuous (bounded or unbounded) in (a, b), $U(x)$ is a lower semi-continuous function, and $L(x)$ is an upper semi-continuous function. Accordingly, $U(x)$ and $L(x)$ are point-wise discontinuous, when they are not continuous.*

If $U(x_1)$ is finite, x_2 can be so determined that

$$U(x_1) - I(x_1, x_2) < \tfrac{1}{2}\epsilon;$$

and also a neighbourhood $(x_1 - \eta, x_1 + \eta)$, of x_1, not containing x_2, can be so determined that, if x_1' be any point in this neighbourhood,

$$|I(x_1', x_2) - I(x_1, x_2)| < \tfrac{1}{2}\epsilon;$$

for $I(x, x_2)$ is continuous with respect to x, at x_1. We have, in this neighbourhood,

$$I(x_1', x_2) > I(x_1, x_2) - \tfrac{1}{2}\epsilon > U(x_1) - \epsilon.$$

It follows that $U(x_1') > U(x_1) - \epsilon$; or $U(x)$ is lower semi-continuous at x_1. If $U(x_1) = +\infty$, x_2 can be so determined that $I(x_1, x_2) > N + \tfrac{1}{2}\epsilon$, where N is an arbitrarily chosen positive number. If $f(x_1)$ is finite, the neighbourhood $(x_1 - \eta, x_1 + \eta)$, of x_1, may be determined as before; and then $I(x_1', x_2) > N$, and thus $U(x_1') > N$, which is the condition that $U(x)$ should be lower semi-continuous at x_1. If $f(x_1) = +\infty$, the neighbourhood can be so determined that $I(x_1', x_2) > N$, and thus $U(x_1') > N$; hence $U(x)$ is lower semi-continuous at x_1. That $L(x)$ is upper semi-continuous can be proved in a similar manner.

The incrementary ratio $I(x, x')$, where $f(x)$ is continuous (bounded or unbounded) in the closed interval (a, b), assumes every value between its upper *and lower boundaries U and L, for a pair of points x, x' interior to the interval (a, b).*

Let k be a number between U and L, thus $U > k > L$. The function $U(x)$ being a lower semi-continuous function, the set of points at which $U(x) \leq k$ is a closed set G_1 (see § 230); and similarly, since $L(x)$ is an upper semi-continuous function, the set at which $L(x) \geq k$ is a closed set G_2.

If, at any point α, $U(\alpha) = L(\alpha) = k$, the function has the value

$$f(\alpha) + k(x - \alpha),$$

for all points x in the interval (a, b). Assume first that no such points α exist: then since $L(x) < U(x)$, the two sets G_1, G_2 can have no points in common. Since the continuum (a, b) cannot be the sum of the two closed sets G_1, G_2, it follows that there exists at least one point which belongs neither to G_1 nor to G_2, and thus at which $U(\alpha) > k > L(\alpha)$. If α be a point at which this condition is satisfied, in a certain neighbourhood of α the conditions

$$U(x) > U(\alpha) - \epsilon, \quad L(x) < L(\alpha) - \epsilon$$

will be satisfied; and, if ϵ be sufficiently small, $U(x) > k > L(x)$, at all points in this neighbourhood. Thus each point α, at which

$$U(\alpha) > k > L(\alpha),$$

is an interior point of a set of points at all of which $U(x) > k > L(x)$. In case there is a point α such that $U(\alpha) = L(\alpha) = k$, and if this does not hold for any smaller value of x than α, it is clear that, in the interval (a, α), the upper and lower boundaries of $I(x)$ are U and L. The above proof applied to (a, α) shews that in this interval there exists a set of points at which $U(x) > k > L(x)$.

At any such point x interior to (a, b), we can choose positive numbers h_1, h_2 such that

$$I(x, x + h_1) > k > I(x, x - h_2),$$

where $x + h_1 < b$, $x + h_2 < b$. Since $I(x, x - h)$ is continuous in the interval of h, of which h_1, h_2 are the end-points, a value of h between h_1 and h_2 exists such that $I(x, x + h) = k$, which establishes the theorem.

If, at every point in (a, b), $f(x)$ have a differential coefficient (finite, or infinite with fixed sign), then $f'(x)$ has every value between its upper and lower boundaries in the interval (a, b).

For, if k be any number between U and L, since, in accordance with the above theorem, a pair of interior points α, β exists, such that

$$I(\alpha, \beta) = k,$$

it follows from the theorem of § 261 that there exists a point x, in (α, β), at which $f'(x) = k$. Since k is any arbitrarily chosen number between U and L, $f'(x)$ takes all values between U and L. It is clear that U and L must be the upper and lower boundaries of $f'(x)$, for if $f'(x) > U$, $x - h$ could be so determined that $I(x, x + h) > U$, which is impossible.

PROPERTIES OF THE DERIVATIVES OF CONTINUOUS FUNCTIONS

279. Let $f(x)$ be continuous in the interval (a, b), and let U and L be the upper and lower boundaries of the incrementary ratios above defined. Take (α, β), any interval in (a, b), and consider the function

$$\phi(x) = f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}[f(\beta) - f(\alpha)].$$

Since $\phi(\alpha) = 0$, $\phi(\beta) = 0$, unless $\phi(x)$ be constant through (α, β), there must be within (α, β) a maximum or minimum of $\phi(x)$; and thus at least one point x_1 exists within (α, β), such that

$$\phi(x_1 \pm h) - \phi(x_1) \leqq 0,$$

for all sufficiently small values of h, or else

$$\phi(x_1 \pm h) - \phi(x_1) \geqq 0,$$

for all sufficiently small values of h. At such a point

$$I(x_1 + h, x_1) \leqq I(\alpha, \beta),$$

and $$I(x_1 - h, x_1) \geqq I(\alpha, \beta),$$

or else $$I(x_1 + h, x_1) \geqq I(\alpha, \beta),$$

and $$I(x_1 - h, x_1) \leqq I(\alpha, \beta).$$

If $\phi(x)$ have an infinite number of maxima and minima in (α, β), there is, in (α, β), an infinite number of points at which the first of the conditions for $\phi(x)$ holds, and also an infinite number at which the second holds. If there be only a finite number of maxima and minima of $\phi(x)$ in (α, β), then this interval can be divided into a number of portions in each of which the function $\phi(x)$ is monotone; and in any one of these portions either

$$I(x \pm h, x) \leqq I(\alpha, \beta),$$

at all points within the sub-interval, or else

$$I(x \pm h, x) \geqq I(\alpha, \beta),$$

for every x within the portion, and for sufficiently small values of h. We see then that $I(\alpha, \beta)$ and $I(x \pm h, x)$ lie between U and L.

Thus, in every interval (α, β) contained in (a, b), in which $\phi(x)$ has an infinite number of maxima and minima, there are (1), an infinity of points x at which $I(x - h, x)$, for all sufficiently small values of h, lies between L and $I(\alpha, \beta)$; and (2), an infinity of points at which the same is true of $I(x - h, x)$: (3), an infinity of points at which $I(x + h, x)$, for all sufficiently small values of h, lies between U and $I(\alpha, \beta)$; and (4), an infinity of points at which the same is true of $I(x - h, x)$.

In case $f(x) - f(\alpha) - \frac{x-\alpha}{\beta-\alpha}[f(\beta) - f(\alpha)]$ have only a finite number of maxima and minima in (α, β), there are in (α, β) finite intervals such that all the points in one of them belong to both the sets (1) and (2), and also finite intervals in which all the points belong to both the sets (3) and (4); each of these sets of intervals is finite, and an interval of one set is followed by one of the other set.

The number L being the lower limit of the function $L(x)$ in the interval (a, b), there exists a point x_1 such that L is the lower limit of the values of $L(x)$ in any arbitrarily small neighbourhood of x_1; and it follows that in such neighbourhood of x_1 there are points ξ such that $I(\xi - h, \xi)$, for an infinity of values of h, differs from L by less than a prescribed positive number ϵ. Therefore there is, in (a, b), an infinity of pairs of points (α, β), one of which is arbitrarily near x_1, such that $I(\alpha, \beta)$ differs from L by less than ϵ.

Similarly, it may be shewn that, in (a, b), there is an infinity of pairs of points (α, β), such that $I(\alpha, \beta)$ differs from U by less than the prescribed number ϵ.

If U or L be infinite, there exists an infinity of pairs of points such that $I(\alpha, \beta)$ is arithmetically greater than a prescribed number c, and has the same sign as the infinite U or L.

We can consequently choose the interval (α, β) so that

$$I(\alpha, \beta) = L + \eta, \text{ or else so that } I(\alpha, \beta) = U - \eta,$$

where $\eta < \epsilon$, provided U and L are finite. If one, or both, of U, L be infinite, (α, β) can be so chosen that $I(\alpha, \beta)$ has the same sign as U, or as L, and is arithmetically greater than a prescribed positive number c.

We have now obtained the following results.

If $f(x)$ be a continuous function, and (a, b) be the whole, or a part, of its domain, to which U and L correspond, then (1), if L be finite, there exists in (a, b) an infinity of points for which $D^-(x)$, $D_-(x)$ both lie between L and $L + \epsilon$, where ϵ is an arbitrarily prescribed positive number; and at these points $D^+(x)$, $D_+(x)$ are either equal, in which case a derivative on the right exists, or else they differ from one another by less than ϵ; (2), if U be finite, there exists in (a, b) an infinity of points for which $D^-(x)$, $D_-(x)$ both lie between U and $U - \epsilon$; and at these points there exist derivatives on the right, or else $D^+(x)$, $D_+(x)$ differ from one another by less than ϵ; (3), if U or L be infinite, there exists an infinity of points at which $D^-(x)$, $D_-(x)$ are both numerically greater than an arbitrarily great number c, and have the same sign as the U or L which is infinite. A similar statement holds as regards the derivatives on the left.

The above is true irrespectively of the number of the maxima and minima of $f(x)$; but if $f(x)$ have in (a, b) only a finite number of maxima and minima, and if the same be true of all the functions $f(x) - lx - m$, obtained by the addition of a linear function, then there exist in (a, b) finite sub-intervals such that at all points in one of them the above statements hold both as regards the derivatives on the right and as regards those on the left. The numbers U and L correspond in each case to the particular sub-interval.

It will be observed that the theorem does not assert the necessity of the existence of points at which a determinate derivative on the right or on the left exists, but it states that there are in every sub-interval points at which the difference between the upper and lower derivatives on one side is less than a prescribed arbitrarily small number, or else at which both such derivatives are arithmetically greater than an arbitrarily fixed large number. There are therefore certainly points in every sub-interval at which there is, so to speak, an arbitrarily near degree of approximation to the existence of a finite or infinite derivative on the right, and also points at which the same is true as regards derivatives on the left.

280. It will now be shewn* that, *for a continuous function, of which* (a, b) *is the whole, or a part, of its domain, the upper boundary of each of the four derivatives* $D^+f(x)$, $D_+f(x)$, $D^-f(x)$, $D_-f(x)$ *for all values of* x *in* (a, b) *is* U, *the upper boundary of the incrementary function in* (a, b), *and that the lower boundary of each of the four functions is* L. *If* U *and* L *be both finite the function belongs to the class of functions with bounded derivatives.*

A function with bounded derivatives accordingly satisfies the condition. that for every pair of points x_1, x_2, $|f(x_1) - f(x_2)| < k\,|x_1 - x_2|$, where k is a fixed positive number. It has been pointed out, in § 275, that such a function belongs to the class of functions of bounded variation.

It is clear that the upper boundary of each of the functions $D^+f(x)$, $D_-f(x)$, $D^-f(x)$, $D_-f(x)$ is a number which cannot be greater than U. Now since it has been shewn that points exist in (a, b) such that, if ϵ be an arbitrarily prescribed number, both $D^+f(x)$, $D_+f(x)$ differ from U by less than ϵ, when U is finite, and are arbitrarily great if U is $+\infty$, it follows that U is in either case the upper boundary of $D^+f(x)$, $D_+f(x)$. In a similar manner it can be shewn that U is the upper boundary of both $D^-f(x)$, $D_-f(x)$. The proof that L is the common lower boundary of the four functions is exactly similar.

Each of the four expressions $D^+f(x)$, $D_-f(x)$, $D^-f(x)$, $D_-f(x)$ may be regarded as a function defined for the whole domain of $f(x)$, except at one of the end-points; the ordinary definition of a function being extended so far as to admit infinite functional values with a fixed sign.

If, at any point x_0, *interior to* (a, b), *one of the above functions, say* $D^-f(x)$, *be continuous, either in the ordinary, or in the extended, sense of the term* (§ 219), *then, at that point, the other three functions are also continuous,*

---

* Du Bois-Reymond, *Math. Ann.* vol. XVI (1880), p. 119; also Scheeffer, *Acta Mathematica*, vol. V (1884), p. 190.

and are equal in value to $D^-f(x_0)$, *and thus there exists, at* x_0, *a differential coefficient.*

To prove this, take any interval $(x_0 - \epsilon, x_0 + \epsilon)$; then all four functions have in this interval the same upper boundaries, and also the same lower boundaries. If $D^+f(x_0)$ be finite, the upper and lower boundaries of $D^-f(x)$ in $(x_0 - \epsilon, x_0 + \epsilon)$ each differ from $D^-f(x_0)$ by less than a number η depending on ϵ, in such a way that, as ϵ is indefinitely diminished to the limit zero, η also diminishes to the limit zero. Since all four functions have the same upper boundary and the same lower boundary in $(x_0 - \epsilon, x_0 + \epsilon)$, the upper and lower boundaries of each differ from $D^-f(x_0)$ by less than η, and η can be made as small as we please by taking ϵ small enough. It follows that all four functions are continuous at x_0, and that all four, at x_0, are equal to $D^-f(x_0)$; and thus there exists a differential coefficient at x_0.

In case $D^+f(x_0)$ is $+\infty$, ϵ can be so chosen that, in $(x_0 - \epsilon, x_0 + \epsilon)$, $D^+f(x)$ is everywhere greater than an arbitrarily large chosen number c, and the upper and lower boundaries of each of the four functions are then greater than c; by taking a succession of values of c which increase indefinitely, and considering the corresponding sequence of values of ϵ which converge to zero, we see that each of the functions $D_+f(x)$, $D^-f(x)$, $D_-f(x)$ is infinite at x_0, and is continuous, in the extended sense of the term, at that point; there is then a differential coefficient at x_0 which is infinite, and of definite sign.

It follows that, *if it be known that any one of the four derivatives is everywhere continuous in an interval, there exists everywhere in the interval a differential coefficient in the ordinary sense of the term.*

281. *The upper and lower boundaries of any one of the four derivatives of a continuous function are the same in the open interval* (a, b) *as in the closed interval**.

It is clear that the upper boundary of $D^-f(x)$ in the open interval cannot exceed that in the closed interval. If possible let it be less by some positive number c. This means that $D^-f(a)$ exceeds the upper boundary in the open interval by c. There exists then no point interior to (a, b) for which $D^-f(x) > D^+f(a) - c$. Now a point x_1 can be so determined that $I(x_1, a)$ differs from $D^-f(a)$ by less than $\frac{1}{2}\epsilon$, and a point x_2 can be so determined that $I(x_1, x_2)$ differs from $I(x_1, a)$ by less than $\frac{1}{2}\epsilon$, and therefore from $D^+f(a)$ by less than ϵ. It follows that, in any interval which includes x_1 and x_2, the upper boundary of $D^-f(x)$ is greater than $D^+f(a) - \epsilon$; and if $\epsilon < c$ this contradicts what has been shewn above. Hence the upper boundary of $D^+f(x)$, in $a \leq x < b$, is the same as that in $a < x < b$.

---

* See W. H. and G. C. Young, *Quarterly Journal of Math.* vol. XL (1909), p. 12.

It follows as a corollary that $D^+f(a)$ cannot exceed the upper limit, on the right, of $D^-f(x)$, at $x = a$.

282. $f(x)$ being continuous in (a, b), the continuous function

$$\phi(x) \equiv f(x) - f(a) - \frac{x-a}{b-a}\{f(b) - f(a)\},$$

which vanishes at a and b, must have a maximum, or a minimum, at an interior point of (a, b). At a maximum α, $D^+\phi(\alpha)$, $D_+\phi(\alpha)$ are both ≤ 0, and $D^-\phi(\alpha)$, $D_-\phi(\alpha)$ are both ≥ 0, the inequalities being both reversed in case α is a minimum. If it be known that, at every interior point of (a, b), there is no distinction of right and left as regards derivatives, so that $D^+\phi(x) = D^-\phi(x)$ and $D_+\phi(x) = D_-\phi(x)$, we see that all four derivatives must vanish at the point α, and thus there is a differential coefficient at α, which has the value 0. We have thus obtained the following extension* of the theorem of the mean:

If there is no distinction of right and left with regard to the derivatives of the function $f(x)$ (bounded or unbounded) continuous in the interval (a, b), there exists a point, interior to the interval, at which a differential coefficient exists equal to $\dfrac{f(b) - f(a)}{b-a}$.

Thus $\dfrac{f(b) - f(a)}{b-a} = f'(a + \theta\overline{b-a})$, where θ is such that $0 < \theta < 1$.

As this may be applied to any interval (α, β) contained in (a, b), we obtain the following theorem:

If there is no distinction, as regards right and left, between the derivatives of the continuous function $f(x)$, defined in the interval (a, b), there exists an everywhere dense set of points in (a, b), of the cardinal number of the continuum, at which $f'(x)$ exists; and $f'(x)$ has every value between its upper and lower boundaries.

283. *The derivatives $D^+f(x)$, $D_+f(x)$ of a continuous function are, at any point x_0, such that $a \leq x_0 < b$, either both continuous on the right, or both of them have a discontinuity of the second kind, on the right; but they cannot have ordinary discontinuities on the right.*

A similar statement holds as regards the continuity or discontinuity of $D^-f(x)$, $D_-f(x)$, on the left.

Suppose that $D^+f(x)$ has, at the point x_0, a limit λ, at x_0, on the right; then, if δ be a prescribed positive number, an interval $(x_0, x_0 + \epsilon)$ can be found, such that $D^+f(x)$, for every point of this interval, except x_0, lies between $\lambda + \delta$ and $\lambda - \delta$. The upper and lower boundaries of each of the four derivatives $D^+f(x)$, $D_+f(x)$, $D^-f(x)$, $D_-f(x)$, for any interval

---

* See W. H. and G. C. Young, *Quarterly Journal of Math.* vol. XL (1909), p. 10.

$(x_0 + \epsilon_1, x_0 + \epsilon)$, where $\epsilon_1 < \epsilon$, must all lie between the values $\lambda + \delta$, $\lambda - \delta$: hence the upper and lower boundaries of $D^+f(x)$, for the interval $(x_0, x_0 + \epsilon)$, lie between these same values, the function $D^+f(x)$ being regarded as undefined at the point x_0; and these upper and lower boundaries of $D^+f(x)$ are the same as those of $D^+f(x)$, $D_+f(x)$, for $(x_0, x_0 + \epsilon)$, the point x_0 being included. It follows that $D^+f(x_0)$, $D_+f(x_0)$, both lie between $\lambda + \delta$ and $\lambda - \delta$; and as this holds for every value of δ, we must have

$$D^+f(x_0) = D_+f(x_0) = \lambda = \lambda';$$

where λ' denotes the limit of $D_+f(x)$, at x_0, on the right; and thus $D^+f(x)$, $D_+f(x)$ are both continuous at x_0, on the right. If $\lambda = +\infty$, then in the interval $(x_0, x_0 + \epsilon)$, at every point except x_0, $D^+f(x) > c$, where c is an arbitrarily chosen number on which ϵ depends; the argument then proceeds as before.

284. *A continuous function $f(x)$ cannot have, at every point of a whole interval, a single-valued derivative on the right, which is everywhere infinite and of the same sign.*

For if $f(x)$ had this property in an interval (α, β), so also would

$$f(x) - f(\alpha) - \frac{x - \alpha}{\beta - \alpha}\{f(\beta) - f(\alpha)\},$$

and this function necessarily has a maximum or minimum within (α, β), which is contrary to the condition that it has a derivative on the right which is always of the same sign, for this involves the condition that the function must constantly increase as x increases from α to β.

Let us now suppose that the continuous function $f(x)$ has, at all points of (a, b), single-valued derivatives on the right (finite or infinite), such that, in a part (α, β), of (a, b), this derivative is continuous at least on one side; *the function $f(x)$ is then such that, at an infinite number of points, it possesses an ordinary differential coefficient.*

The derivative $Df(x)$, on the right, cannot at all points of (α, β) be infinite. For if we take a point x_0 such that it is continuous on one side, in the extended sense of the term explained in § 219, then if it were everywhere infinite, its sign at all points in an interval on the one side of x_0 would be the same; but it has been shewn to be impossible that, everywhere in any interval, $Df(x)$ should be infinite and of constant sign. It follows that there are points in the neighbourhood of x_0 at which $Df(x)$ is finite. If x_1 be such a point in (α, β), then, since $Df(x)$ is continuous on one side at x_1, an interval can be found at all points of which it is finite, and also continuous on one side. If (α_1, β_1) be such an interval in (α, β), then since $Df(x)$ is everywhere finite in it, and continuous on one side at least, it is a point-wise discontinuous function, if it be not continuous in (α_1, β_1); and there must therefore be an infinity of points in (α_1, β_1) at which $Df(x)$ is continuous. At such points, in accordance with § 280, $f(x)$ has a differential coefficient.

285. As regards functions which have an infinite number of oscillations in the neighbourhood of a particular point, the following remarks may be made.

If a continuous function have, in every neighbourhood on the right of a point x_0, an infinite number of maxima and minima, there is, in such neighbourhoods, an infinity of points at which the derivatives on the right are negative or zero, and an infinity of points at which these derivatives are positive or zero. It follows from this, that none of the derivatives at a point x, on the right of x_0, can have a definite limit, as x approaches the limit x_0, unless such limit be zero.

In particular, if at all such points x, definite derivatives on the right and on the left exist, these derivatives cannot be continuous at x_0, unless the derivatives at x_0 are both zero.

If, at the point x_0, and at every point in the neighbourhood of x_0 which contains an infinite number of maxima and minima, a differential coefficient exist, which is continuous at x_0, this differential coefficient must be zero at x_0 and at an infinity of points in the neighbourhood of x_0, and must therefore itself have an infinite number of maxima and minima in the neighbourhood of x_0.

If a function $f(x)$, which has an infinite number of maxima and minima in the neighbourhood of x_0, have at x_0, and in its neighbourhood, differential coefficients of any number of orders, then they are all functions with an infinite number of maxima and minima in the neighbourhood of x_0, and all of them vanish at x_0, except that the one of highest order may be discontinuous at x_0, not then necessarily vanishing at that point. If differential coefficients of all orders exist, they must all vanish at x_0; and such a function is incapable of expansion in powers of $x - x_0$ in the neighbourhood of x_0. An example* of a function of this kind is

$$x^2 + e^{-\frac{1}{(x-x_0)^2}} \sin \frac{1}{x - x_0}.$$

FUNCTIONS WITH ONE DERIVATIVE ASSIGNED

286. *If two functions, defined for a given interval, have each bounded derivatives, and if the two functions have each a particular one of their four derivatives, say the upper derivative on the right, equal to one another at every point which does not belong to a set of points E, of measure zero, then the two functions differ from one another by a constant, the same for the whole interval.*

This theorem† differs from that of § 267, in that the functions are restricted to be such continuous functions as have bounded derivatives; it is however more general, in that E is not restricted to be enumerable.

---

* Dini-Lüroth, *Grundlagen*, p. 314.

† Lebesgue's *Leçons sur l'intégration* (1904), p. 79.

Let the points of E be enclosed in the interiors of intervals of a set, of which the total length has the arbitrarily small value ϵ. To each point P, of E, there corresponds an interval PP', where PP' is that part of the interval of the set that encloses P which is on the right of P; these intervals PP' may be denoted by δ'. If the two functions $f_1(x)$, $f_2(x)$ are such that, at a point x_1, $D^+f_1(x_1) = D^+f_2(x_1)$, it has been seen in § 267 that

$$D^+f(x_1) \geq 0, \quad D_-f(x_1) \leq 0,$$

where $f(x)$ denotes $f_1(x) - f_2(x)$. Since $f(x)$ is continuous at x_1, it follows that there is a set of points $x_1 - h$, on the right of x_1, such that

$$|f(x_1 - h) - f(x_1)| \leq \epsilon h;$$

if we suppose h to have the greatest value for which this holds, the interval $(x_1, x_1 + h)$ is an interval on the right of x_1, and such intervals may be denoted by δ.

Let ξ be any point such that $a < \xi < b$, and consider the interval (a, ξ). From the point a lay off an interval δ, or δ', according as a is not, or is, a point of E; from the end of this interval lay off another interval δ, or δ', as the case may be. Proceeding in this manner, we may construct a Lebesgue chain reaching from a to ξ (see § 78). The set of points not interior to the intervals of the chain is a closed enumerable set. We can now find $f(\xi) - f(a)$ as the sum, or limiting sum, of the differences of the functional values at the end-points of the intervals of the chain, and each of which is either a δ, or a δ'. It is clear that

$$|f(\xi) - f(a)| \leq \epsilon\Sigma\delta + A\Sigma\delta' < \epsilon(\xi - a - A),$$

where the summations refer to those of the intervals δ, δ' which have been employed in the construction, and A denotes the finite upper boundary of $|I(x_1, x_2)|$, for every pair of points, x_1, x_2 in the interval (a, ξ); and this is identical with the upper boundary of the absolute value of the derivatives of $f(x)$ in the interval. Since ϵ is arbitrarily small, it follows that $f(\xi) = f(a)$, and therefore $f_1(\xi) - f_2(\xi) = f_1(a) - f_2(a)$; thus the theorem has been established.

THE CONSTRUCTION OF CONTINUOUS FUNCTIONS

287. One of the most fruitful methods of obtaining continuous functions which exhibit various peculiarities as regards the existence or non-existence of differential coefficients at all the points, or at sets of points, of their domain, consists in defining the functions by means of series specially constructed with a view to the purpose in hand: this method will be explained and illustrated in Vol. II. Brodén, Köpcke, and others, have however given direct constructions for continuous functions, which illustrate various possibilities in relation to the existence and properties of derivatives.

The method employed* by Brodén is that of defining a continuous function in the domain (a, b), as the function obtained by *extension* (see § 241) of a function defined for an enumerable everywhere dense set of points in (a, b), the primary points. A continuous function is entirely determinate when the functional values at such a primary set of points have been assigned. The necessary and sufficient condition that a function defined for the primary set should, by extension to the domain (a, b), give a function which is continuous in that domain, is that the primary function should be *uniformly continuous* with respect to the unclosed primary domain. To prove this, let $\{\xi\}$ denote the set of primary points, and $\{x\}$ the set of secondary points; then the condition that the function $f(\xi)$ may be uniformly continuous with respect to the domain $\{\xi\}$, is that, if ξ_1 be any point of $\{\xi\}$, and if η be a prescribed arbitrarily small number, the condition $|f(\xi) - f(\xi_1)| < \eta$ be satisfied at all points ξ which are such that $|\xi - \xi_1| < \epsilon$, where ϵ is a number dependent on η, the same for all points ξ_1 of $\{\xi\}$. Now, assuming that this condition is satisfied, let x_1 be a secondary point, and let $\xi_1, \xi_2, \dots \xi_n \dots$, $\xi_1', \xi_2', \dots \xi_n' \dots$ be any two sequences of primary points, each of which has x_1 as its limit; we have to shew that each of the sequences

$$f(\xi_1), f(\xi_2), \dots f(\xi_n) \dots,$$
$$f(\xi_1'), f(\xi_2'), \dots f(\xi_n') \dots$$

converges to the same number, which will then be the single functional value $f(x_1)$. Enclose x_1 in the interval $(x_1 - \frac{1}{2}\epsilon, x_1 + \frac{1}{2}\epsilon)$; then, from and after some particular value of n, all the points of both sequences of values of ξ lie within this neighbourhood. Let this value of n be m, then

$$|f(\xi_m) - f(\xi_{m+r})| < \eta,$$

for all positive integral values of r; hence the first sequence of functional values is convergent, since η is arbitrary; and similarly the second is also convergent. Also for every η there is a definite m such that

$$|f(\xi_{m+r}) - f(\xi'_{m+r})| < \eta;$$

hence the two convergent sequences have the same limit, and this limit defines $f(x_1)$. We have now to shew that the single-valued function so defined is continuous. We have

$$|f(x_1) - f(\xi_1)| < \eta, \text{ provided } |x_1 - \xi_1| < \tfrac{1}{2}\epsilon,$$

and $$|f(x_2) - f(\xi_2)| < \eta, \text{ provided } |x_2 - \xi_2| < \tfrac{1}{2}\epsilon;$$

also $$|f(\xi_2) - f(\xi_1)| < \eta, \text{ provided } |\xi_2 - \xi_1| < \epsilon.$$

Hence it follows that $$|f(x_2) - f(x_1)| < 3\eta,$$

and this holds provided $$|x_2 - x_1| < 2\epsilon,$$

---

* *Crelle's Journal*, vol. CXVIII (1897), p. 1. See also *Acta Univ. Lund.* vol. XXXIII (1897); *Functiontheoretische Bemerkungen und Sätze.*

for ξ_1, ξ_2 can be taken to be between x_1 and x_2: and therefore $f(x)$ is continuous at x_1, since 3η is at our choice. The extended $f(x)$ is also easily seen to be continuous at any primary point. It has now been proved that the condition of uniform continuity is sufficient; that it is necessary follows from the theorem of § 217.

The derivatives at any point depend only on the functional values at the primary points in the neighbourhood of the point. For let x_1 be any point, and consider the limit of $\dfrac{f(x)-f(x_1)}{x-x_1}$, when x has any sequence of values which converge to x_1. A set of primary values of x can always be found, such that the ratio converges to the same limit, when x has the values of this sequence of primary points, as for the prescribed sequence consisting of secondary points, or of both primary and secondary points. For a primary point ξ can be found, corresponding to x, such that

$$\left|\frac{f(x)-f(x_1)}{x-x_1}-\frac{f(\xi)-f(x_1)}{\xi-x_1}\right|<\delta,$$

where δ is an arbitrarily small number. This follows from the fact that

$$\frac{f(x)-f(x_1)}{x-x_1}$$

is a continuous function of x at every point except x_1.

288. In order to construct monotone continuous functions the values of the function are first assigned at the end-points a, b of the interval, then at two points x_0, x_1, where $x_0 < x_1$, then at four points x_{00}, x_{01}, x_{10}, x_{11}, where

$$x_{00} < x_0, \quad x_0 < x_{01} < x_{10} < x_1, \quad \text{and} \quad x_1 < x_{11};$$

afterwards at eight points

$$x_{000},\ x_{001},\ x_{010},\ x_{011},\ x_{100},\ x_{101},\ x_{110},\ x_{111},\ \&c.$$

lying in the successive intervals measured from left to right, into which (a, b) was divided by the four points: and so on. The function may then be regarded as the limit of a sequence of continuous functions, each of which is representable as a polygon obtained by joining the end-points of ordinates which represent the functional values that have been assigned at any stage of the process.

In this manner Brodén has constructed a monotone continuous function $f(x)$, which is such that it has derivatives on the right, and on the left, which are everywhere definite, finite, and different from zero; and such that a definite differential coefficient everywhere exists, except at the everywhere dense enumerable set of primary points.

He has also constructed a monotone function $f(x)$ which is such that, at the everywhere dense enumerable set of primary points, the derivative on the left exists, and is zero, and the derivative on the right exists, and is positive: for an unenumerable everywhere dense set of points there is a differential coefficient, everywhere zero: and for another such set of points, there is no definite derivative on the left, but there is a positive one on the right.

A third case is the following:

$f(x)$ is continuous, monotone, and increasing; at an everywhere dense enumerable set of points the derivative on the left is zero, and that on the right is $+\infty$; for an everywhere dense unenumerable set, both derivatives exist, and are positive; for another such set both exist, and are zero: for a third such set, both derivatives exist, and are $+\infty$; for a fourth such set, neither derivative exists; for a fifth such set, the derivative on the left is zero, and that on the right is indefinite, but has zero for its lower limit: for a sixth such set, the derivative on the right is $+\infty$, but that on the left is indefinite, with $+\infty$ for its upper limit.

289. For the construction of everywhere oscillating continuous functions it is more convenient to assign successively the functional values at sets of points proceeding by powers of 3, instead of 2, as in the case of monotone functions. In this manner Brodén has constructed such a function $f(x)$, which has the following properties:

At an everywhere dense enumerable set of points, the derivative on the left exists, and is positive: that on the right exists, and is negative (or the reverse), this set corresponding to maxima and minima of the function: for a certain unenumerable everywhere dense set, there is a differential coefficient everywhere of the same sign: and for another such set, there is a differential coefficient which is zero; for a third such set, one or both of the derivatives are indefinite.

Köpcke[*] has given the first example of a function which is everywhere oscillating, and yet has at every point a definite differential coefficient, thus confirming the conjecture of Dini that such functions can exist; and Brodén[†] has also constructed such a function. A general theory of such functions has been given[‡] by Schoenflies. The method adopted by Köpcke is to construct the function as the limit of a succession of polygons of which the sides are circular arcs. Everywhere oscillating functions have also been studied by Steinitz[§]. A detailed account of all the special cases treated of by these writers would require much space; reference can therefore only be made to the original memoirs. A simplification of Köpcke's construction, due to Pereno, will be given in Vol. II.

---

[*] *Math. Ann.* vol. XXIX (1887), p. 123; vol. XXXIV (1889), p. 161; vol. XXXV (1890), p. 104. See also Pereno, *Giorn. di Mat.* vol. XXXV (1897), p. 132.

[†] *Stockholm Vet. Ak. Öfv.* (1900), pp. 423 and 743.

[‡] *Math. Ann.* vol. LIV (1901), p. 553: also his *Bericht*, vol. I, p. 164.

[§] *Math. Annalen*, vol. LII (1899), p. 58.

290. A function $f(x)$ which is of such a character that it can be represented approximately by a graph that exhibits all the peculiarities of the function, so that $y = f(x)$ is the equation of a "curve," in the ordinary sense of the term, must satisfy the following three conditions:

(1). The function must be continuous everywhere, with the possible exception of a finite number of points, at which it may have ordinary discontinuities.

(2). It must be differentiable, except that there may be a finite number of points at which no differential coefficients exist, but at which definite derivatives on the right and on the left exist.

(3). It can have only a finite number of maxima and minima; and the same must hold of every function obtainable by the addition of a linear function to the one in question. This condition may be expressed in the form, that the function must be in general monotone with reference to every possible axis which may be employed for the measurement of abscissae.

A function which satisfies these conditions may be characterized as an *ordinary* function. As has been already indicated, there exist functions which satisfy the conditions (1) and (3), but do not satisfy the condition (2). Again, there exist functions which satisfy the conditions (1) and (2), but not the condition (3).

GENERAL PROPERTIES OF DERIVATIVES

291. A number of important properties of the derivatives of a continuous function have been investigated in §§ 279–285. Some of the most interesting properties of the derivatives of a function hold for functions that are not restricted to be continuous, and are either quite general or are only restricted to belong to the wide class of measurable functions. It will in general be assumed that a function $f(x)$, defined for any interval of x, either finite or indefinitely great, has a definite finite value for each value of x, whether the function be bounded or not.

If $f(x)$ be defined in a finite, or an unbounded, interval, and k be any fixed number, the set of points x, at each of which the conditions $D_+ f(x) \geq k$, $D^- f(x) < k$ are satisfied, is enumerable[†].

Let G be the set of points at which $D_+ f(x) \geq k$; and let $\{\epsilon_r\}$, $\{\eta_r\}$ be two monotone sequences of positive numbers that converge to zero. If ξ be a point of G, and h be a sufficiently small positive number,

$$I(\xi, \xi + h) \geq k - \eta_n.$$

---

[*] Du Bois-Reymond, *Crelle's Journal*, vol. LXXIX (1875), p. 32.

[†] See G. C. Young, *Acta Math.* vol. XXXVII (1914), p. 143, and *Quarterly Journal of Math.* vol. XLVII (1916), p. 127.

All the values of h for which this condition is satisfied have an upper boundary $\bar{h}$, at which the condition may, or may not, be satisfied. Let $\bar{\xi} = \xi - \frac{1}{2}\bar{h}$, then we have, for such a point ξ, of G, a definite interval $(\xi, \bar{\xi})$, at every point of which the condition $I(\xi, x) \geq k - \eta_n$ is satisfied. We now take, for the point ξ, the interval δ_ξ on its right, of length equal to the smaller of the two numbers ϵ_n, $\bar{\xi} - \xi$, so that, in this interval, both the conditions

$$I(\xi, x) \geq k - \eta_n, \quad |\xi - x| \leq \epsilon_n$$

are satisfied. Let Δ_n be the set of all such intervals δ_ξ, when every point ξ, in G, is considered; and let Δ be the set $\{\Delta_n\}$ of all such intervals, when all values of n are taken into account. There exists at most an enumerable set H, of the points of G, that are not interior points of any interval of Δ (see § 72). If ξ' be a point of $G - H$, it is interior to an interval (ξ, η), of Δ_n, whatever value of n is taken. It follows that $I(\xi, \xi') \geq k - \eta_n$. As n is indefinitely increased, ξ will converge to ξ', and η_n converges to 0; it is thus seen that $D^-f(\xi') \geq k$. This result holds at every point ξ', of G, that does not belong to the enumerable set H, and thus the theorem has been established.

It is clear that the corresponding result holds, that the set of points at which $D_-f(x) \geq k$, $D^-f(x) < k$ is enumerable.

If $f(x)$ be changed into $-f(x)$, and k into $-k$, we see that the set of points at which $D^-f(x) \leq k$, $D_-f(x) > k$ is enumerable, and that the set of points at which $D^-f(x) \leq k$, $D_-f(x) > k$ is also enumerable.

It is clear that the set of points, at each of which $D_+f(x) = +\infty$, $D^-f(x) < k$, being a part of the enumerable set at which $D_+f(x) \geq k$, $D^-f(x) < k$, is enumerable. If we give to k successively the values in a divergent sequence $\{k_n\}$ of positive numbers, the set at which

$$D_-f(x) = +\infty, \quad D^-f(x) < \infty,$$

is such that each point belongs to one of the enumerable sets for which $D_-f(x) = +\infty$, $D^-f(x) < k_n$; it has thus been shewn that:

The set of points at which $D_-f(x) = +\infty$, and $D^-f(x)$ is finite, or $-\infty$, is enumerable.

Similarly the three sets at which $D_-f(x)$ is $+\infty$, and $D^+f(x)$ is finite, or $-\infty$: at which $D^+f(x)$ is $-\infty$, and $D_-f(x)$ is finite or $+\infty$; and at which $D^-f(x)$ is $-\infty$, and $D_-f(x)$ is finite or $+\infty$, are all enumerable.

292. *Except at points of an enumerable set the upper derivative on one side is greater than, or equal to, the lower derivative on the other side.*

Let us consider a point x, at which $D^+f(x) = a$, $D_-f(x) = b$, where a and b are numbers such that $a < b$; the number b may be $+\infty$. Let k be a rational number interior to the interval (a, b), then x belongs to the set of points at which $D^+f(x) < k$, $D_-f(x) > k$, and this set E_k, being a part of the enumerable set at which $D^+f(x) < k$, $D_-f(x) > k$, is enumerable. Any point x, at which $D^+f(x) < D_-f(x)$, belongs to E_k for all properly chosen values of k. Any point at which $D^+f(x)$ is finite, or $-\infty$, and also $D_-f(x) = +\infty$, belongs to the enumerable set E_∞. All the sets E_k for all rational values of k, and also for $k = -\infty$, form an enumerable set. Therefore all the points at which $D^+f(x) < D_-f(x)$ form an enumerable set. Similarly, it is seen that the set of points at which $D^-f(x) < D_+f(x)$ is enumerable, and thus the theorem is established.

In the case of a continuous function $f(x)$, there is a derivative on one side which has any assigned value not greater than the upper derivative, and not less than the lower derivative on that side (see § 290). At every point x that does not belong to a certain enumerable set,

$$D^+f(x) \geq D_-f(x), \text{ and } D^-f(x) \geq D_+f(x);$$

thus at least one point of the closed interval $(D_-f(x), D^+f(x))$ lies in the closed interval $(D_+f(x), D^-f(x))$. We have then the following theorem:

A continuous function has at least one symmetrical median, or extreme, derivative at every point that does not belong to a certain set of measure zero.

The following theorem follows at once from the above general theorem:

Except at points belonging to an enumerable set, a function cannot have unique progressive and regressive derivatives which are unequal, whether finite, or infinite with a determinate sign.

293. The following theorem*, which will be of use in the further discussion, will now be established:

If $f(x)$ is continuous with respect to a perfect set G, contained in an interval (a, b), in which $f(x)$ is defined, and if, at each point x, of G, there is a closed interval $(x, x + h_x)$ on its right, such that, if x' be any point in it, other than x itself, $f(x') \leq f(x)$, then an interval (α, β) exists, in (a, b), containing a part of G, such that, for every pair of points x, x', in (α, β), which are such that x belongs to G, and $x' \geq x$, the condition $f(x') \leq f(x)$ is satisfied.

Let a system of nets with closed meshes be fitted on to the interval (a, b). If, in any net D_n, there is a mesh d_n which contains, in its interior, points of G such that, for each such point x, the condition $f(x') \leq f(x)$ is satisfied for all points x', in the mesh d_n, which are on the right of x, the mesh d_n, or else that part of it obtained by cutting off a portion of its left end, will be an interval such as is required. Let it be assumed that, if possible, for each value of n, and for every mesh d_n, which contains points of G in its interior, there is at least one pair of points x, x' ($> x$), where x is a point of G, and such that $f(x') > f(x)$; and where x is interior to d_n.

---

* See G. C. Young, *Proc. Lond. Math. Soc.* (2), vol. xv (1916), p. 367; the proof there given differs from that in the text, in that it seems to require an infinite number of undetermined acts of choice, as applied to successive pairs of points. For corrections, see *Proc. Lond. Math. Soc.* (2), vol. xix (1921), p. 152.

In D_n, let H_n be the set of those points x, of G, each of which is interior to a mesh of D_n, and such that there is a point x' ($> x$) in the same mesh, for which $f(x') > f(x)$. Let K_n be the set of all other points of G; thus $G = H_n + K_n$.

It will be shewn that, on the assumption that the theorem does not hold, the set K_n is non-dense in G. Let Δ be any interval that contains points of G in its interior. For a sufficiently large value of n, Δ will contain, in its interior, a mesh of D_n that has, within it, points of G; one of such points $\bar{x}$ is such that $f(x') > f(\bar{x})$, for some point x' ($> \bar{x}$) contained in the mesh. A neighbourhood $(\bar{x} - \epsilon, \bar{x} + \epsilon')$ of $\bar{x}$, contained in the mesh, can be so determined that, for all points x, of G, in that interval, $f(x') > f(x)$; this follows from the continuity of $f(x)$ in G. Therefore no point ξ, of G, in $(\bar{x} - \epsilon, \bar{x} + \epsilon')$ is a point of K_n. Since, interior to an arbitrary interval Δ which contains in its interior points of G, another interval $(\bar{x} - \epsilon, \bar{x} + \epsilon')$ can be so determined as to contain points of G, none of which belong to K_n, it follows that K_n is non-dense in G.

We see that H_n contains H_{n-1}; and the sets K_n being all non-dense in G, the set $M(K_1, K_2, \ldots K_n, \ldots)$ is of the first category relatively to G. It follows that there exists a set H_ω, a residual with respect to G, that consists of points each of which belongs to H_n, for every value of n. If x be a point of H_ω, there exists an interval $x + h_x$, such that, for every point x' in it, $f(x') \leq f(x)$.

If n be sufficiently large, the part of that mesh of D_n containing x, that is on the right of x, will be contained in $(x, x + h_x)$, and consequently, in that mesh there is no point x' ($> x$) such that $f(x') > f(x)$; and this is contrary to the hypothesis. There is consequently a contradiction in the assumption made. Hence the theorem has been established.

294. The following theorem can be deduced from that of § 293:

If, at each point of the perfect set G, with respect to which $f(x)$ is continuous, $D^- f(x) < 0$, an interval (α, β) can be so determined that, for any interval contained in it, with a point of G as left-hand end-point, the incrementary ratio is ≤ 0.

For if, at a point x, of G, we have $D^- f(x) < 0$, a set of intervals $(x, x + h)$ exists, such that $f(\xi) \leq f(x)$, for all points ξ ($\neq x$) of $(x, x + h)$. There will be an upper boundary $\bar{h}$, of h, and at the point $x + \bar{h}$, we may have $f(x + \bar{h}) > f(x)$. The neighbourhood $(x, x + \frac{1}{2}\bar{h})$ may be assigned to x as the interval $(x, x + h_x)$ in the theorem proved above. In the interval (α, β), we have $I(x, x') \leq 0$, at every point x, of G.

If we replace $f(x)$, in the above theorem, by $f(x) - kx$, we have the following theorem:

If, at each point of the perfect set G, with respect to which $f(x)$ is continuous, $D^+f(x) < k$, an interval (α, β) can be so determined that, for any interval contained in it, with a point of G as left-hand end-point, the incrementary ratio is $\leqq k$.

If we change $f(x)$ into $-f(x)$, and k into $-k$; since

$$D^+\{-f(x)\} = -D_+f(x)$$

we have the corresponding theorem, in which $D_+f(x) > k$ takes the place of the condition $D^+f(x) < k$, and the incrementary ratio in (α, β) is $\geqq k$.

It is clear that corresponding theorems hold for $D^-f(x)$, $D_-f(x)$.

295. In order not to interrupt the continuity of the account of the main properties of the derivatives of functions, and to secure the greatest attainable degree of generality in the results, the conception of a measurable function will be here introduced. A function $f(x)$, defined in any interval (a, b), is said to be *measurable*, provided that, for every value of A, the set of points x, of (a, b), at which $f(x) > A$, is a measurable set of points. The properties of measurable functions will be considered in detail in the account of Lebesgue integrals which will be given in Chapter VII. An important property of a measurable function $f(x)$ will here be assumed, viz. that a set E, of points of (a, b), of measure arbitrarily near to $b - a$, the measure of the interval, exists such that $f(x)$ is continuous with respect to the set E. This property of measurable functions will be established in Vol. II. If this property of a measurable function be not assumed, the properties of derivatives established below must be taken only to apply to continuous functions. It has been shewn in § 128, that any measurable set E contains a perfect set G, such that $m(E) - m(G)$ is less than an arbitrarily chosen positive number, it being assumed that $m(E) > 0$. It has further been shewn, in § 138, that a perfect set G, of measure greater than zero, contains a perfect component $\bar{G}$, of measure equal to that of G, such that $\bar{G}$ is metrically dense in itself. It follows that a measurable function $f(x)$, defined in the interval (a, b), is continuous with respect to a perfect set $\bar{G}$, metrically dense in itself, and such that $m(\bar{G})$ is less than $b - a$, by less than an arbitrarily chosen positive number.

If H be a measurable set of points contained in (a, b), and of measure $m(H) > 0$, the set H must contain a perfect component, metrically dense in itself, and such that the measurable function $f(x)$ is continuous with respect to that perfect component. For $\bar{G}$ may be so defined that

$$b - a - m(\bar{G}) < m(H),$$

and thus H and $\bar{G}$ have a part K in common, such that $m(K) > 0$. The set K contains a perfect component, metrically dense in itself, and this is the required component of H, with respect to which $f(x)$ is continuous.

296. *The set of points at which, for a finite measurable function* $f(x)$, *the conditions* $D^+f(x) = +\infty$, $D_-f(x) > 0$ *hold, forms a set of measure zero.*

A function is said to be finite, whether it be bounded or not, when its value at no point of its domain is infinite.

Suppose, if possible, that the set had a measure > 0. A part G of this set can be taken, which is perfect, and with respect to which $f(x)$ is continuous, and such that G is metrically dense in itself. An interval (a, b) can be so determined (see § 294) that, if ξ be any point of it belonging to G, $\dfrac{f(\xi) - f(x)}{\xi - x} \geqq 0$, if $a \leqq x < \xi$. Let E be the part of G in (a, b); then $m(E) > 0$.

To each point of (a, b) an interval, with the point as left-hand end-point, can be assigned as follows:

To each point of E, we assign an interval such that the incrementary ratio in it is $\geqq A$, a fixed number. To each point in an interval contiguous to E, we assign that part of the contiguous interval which is on its right. To b, an arbitrary interval can be assigned. One Lebesgue chain can be defined, constituted of intervals of the set, which reaches from a to b; we can then omit the interval corresponding to b.

The sum of all those intervals of the chain of which the left-hand end-point belongs to E has measure $> m(E)$; for these intervals contain all the points of E, and the other intervals contain no points of E in their interiors.

A finite set Δ, of intervals of the chain, can be so chosen that the remainder has measure $< \frac{1}{2}m(E)$.

The finite set contains therefore intervals Δ', of which the left-hand end-points belong to E, and the sum of such intervals is $> \frac{1}{2}m(E)$. In the intervals of (a, b) complementary to Δ', the incrementary ratio is $\geqq 0$. In Δ' and its complement the increment of $f(x)$ from left to right is $\geqq 0$; and the total of all such increments is $f(b) - f(a)$. But these intervals have been shewn to contain a set in which the sum of the increments is $> \frac{1}{2}Am(E)$, which exceeds $f(b) - f(a)$, provided A is chosen sufficiently large; and this is impossible. Therefore the measure of the set is 0.

If we substitute $f(x) - kx$ for $f(x)$, we see that the set of points at which $D^+f(x) = +\infty$, $D_-f(x) > -k$, is zero.

Giving to k successively, the values in a set $k_1, k_2, \ldots$ which increase indefinitely, such that $k_n < k_{n-1}$, each of the sets corresponding to $k_1, k_2, \ldots$, has measure zero, hence the set at each point of which $D^+f(x) = +\infty$, $D_-f(x) > -k_n$, for some value of n, has measure zero.

Therefore, at every point at which $D^{-}f(x) = -\infty$, with the exception of a set of measure zero, we must have $D_{-}f(x) = -\infty$.

Hence we have the following theorems:

(1). *If $f(x)$ be a finite measurable function, and if, at the points of a set, $D^{-}f(x) = +\infty$, then $D_{-}f(x) = -\infty$ at all the points of that set, with the exception of a part, of which the measure is zero. In the set in which $D_{+}f(x) = -\infty$, with a similar exception, $D^{-}f(x) = -\infty$.*

(2). *If $f(x)$ be a finite measurable function, the set of points at which it has an infinite unique derivative, on one side, has the measure zero.*

Let $D^{-}f(x) = D_{-}f(x) = -\infty$, at all points of a set G. Since

$$D^{-}f(x) \geqq D_{-}f(x),$$

except at points of a set of measure zero, we must have $D^{-}f(x) = +\infty$, if the exceptional set be left out of account. By the last theorem, when $D^{-}f(x) = +\infty$, we must have $D_{-}f(x) = -\infty$, except at a set of measure zero. It follows that the set G has measure zero.

297. (3). *If $f(x)$ is a measurable and finite function, the points at which $D^{+}f(x)$, $D_{-}f(x)$ are finite, and different from one another, form a set of measure zero.*

If S is the set of such points, and we remove from S some set of measure zero, the derivatives $D^{-}f(x)$, $D_{-}f(x)$ must also be finite. This follows from theorem (1). We may assume then that S is a set in which all the derivatives are finite at each point, and $D^{+}f(x) \neq D_{-}f(x)$. Let us suppose that, if possible $m(S) > 0$. If N be a positive number, let S_N be that part of S in which all the four derivatives are numerically $< N$. If N have the values in an increasing sequence $N_1, N_2, \ldots$ which diverges, we see that $S_{N_{r+1}}$ contains S_{N_r}; the set S is the outer limiting set of the sequence $\{S_{N_r}\}$, and since $m(S) > 0$, we must have $m(S_r) > 0$, from and after some value of r. We may then take N so large that $m(S_N) > 0$

A part Σ, of S_N, can be determined so as to be perfect, and such that the measure of that part of it that lies in any interval which contains, in its interior, points of Σ, has measure > 0; and is also such that $f(x)$ is continuous relatively to Σ (see § 295).

Let $g(x) = f(x) + Nx$, then all the derivatives $D^{-}g(x)$, $D_{-}g(x)$, $D^{-}g(x)$, $D_{-}g(x)$ are, in Σ, $< 2N$, and > 0.

An interval (a, b) can be chosen such that it contains a part G, of Σ, such that the incrementary ratio of $g(x)$, for any pair of points contained in the interval, is $\leqq 2N$, if the right-hand point belongs to G; or is $\geqq 0$, if the left-hand point belongs to G. The interval can be so chosen that a and b are points of G.

A part G_k, of G, can be so chosen that $D^{+}g(x) - D_{-}g(x) > k$, and k can be so chosen that $m(G_k) > 0$.

For, if k have the values of a decreasing sequence of numbers that converges to 0, G_{k_r} is contained in $G_{k_{r+1}}$; thus the measure of G_{k_r}, from and after some value of r, must be > 0. The set G_k is the sum of a number of sets G_{k_ι}, where ι is an integer, such that, in G_{k_ι},

$$\tfrac{1}{2}(\iota - 1)\, k \leq D_+ g(x) < \tfrac{1}{2}\iota k,$$

where $\iota = 1, 2, 3, \dots \iota_1$; ι_1 being the greatest integer for which

$$\tfrac{1}{2}(\iota - 1)\, k < 2N.$$

One at least of these sets G_{k_ι} must have measure > 0; let ι be the least integer for which this is the case. There exists a part E, of G_{k_ι}, that is perfect and has measure > 0. We may, by alteration of a and b, if necessary, ensure that a and b belong to E. In the set E,

$$\tfrac{1}{2}(\iota - 1)\, k \leq D_+ g(x) < \tfrac{1}{2}\iota k, \text{ and } D^+ g(x) > \tfrac{1}{2}(\iota + 1)\, k.$$

Choose a finite set of the intervals that are contiguous to E, such that the measure of the sum of the remaining intervals is $< \epsilon$, where

$$0 < \epsilon < m(E).$$

Let this finite set be denoted by Δ', and let Δ be the finite set complementary to Δ', in (a, b). We have then

$$m(\Delta) > m(E), \text{ and } m(\Delta) < m(E) + \epsilon.$$

Let us consider one of the intervals (α, β), of Δ; α, β are points of E. To each point x, of the part of E interior to (α, β), and to the point α, we assign an interval $(x, x + h)$, on its right, such that

$$g(x + h) - g(x) < \tfrac{1}{2}\iota k h.$$

To each point in (α, β), not a point of E, we assign that part of the interval contiguous to E in which it lies, that is on its right. There is a unique chain composed of intervals so defined, that stretches from α to β. Remove from this chain a set of intervals of total measure $< \epsilon/n$, such that the remaining intervals $\Delta''_{\alpha,\beta}$ form a finite set. The sum of the finite set of intervals complementary to $\Delta''_{\alpha,\beta}$ is a finite set $\Delta'''_{\alpha,\beta}$, of total measure $< \epsilon/n$, where n is the number of intervals (α, β), of Δ. We proceed in this way with each interval (α, β), of Δ.

The interval (a, b) is now divided into a finite number of parts consisting of (1), the set $\Delta'' \equiv \Sigma\Delta''_{\alpha,\beta}$, (2), the set $\Delta''' \equiv \Sigma\Delta'''_{\alpha,\beta}$, and (3), the set Δ'.

We have $m(\Delta''') < \epsilon$.

The intervals of Δ'' consist partly of intervals $(x, x + h)$, where x is a point of E, and partly of intervals not containing in their interiors points of E. The measure of the latter set is $< \epsilon$, and that of the former is $> m(E) - \epsilon$, and $< m(E) + \epsilon$.

Since, for one of these intervals $(x, x + h)$, we have

$$g(x + h) - g(x) < \tfrac{1}{2}\iota k h,$$

and in any interval with the left-hand end-point a point of E, the incrementary ratio is $\leqq 2N$; we have

$$g(b) - g(a) < \tfrac{1}{2}\iota k \{m(E) - \epsilon\} - 2N\epsilon - p, \quad \text{.........}(1)$$

where p denotes the sum of the increments of $g(x)$ in the intervals Δ'.

Again, if we assigned to each point x, of E, an interval $(x, x + h')$ such that $g(x + h') - g(x) > \frac{1}{2}(\iota + 1) k$, and proceeded as before, then, remembering that, in any interval in which the left-hand end-point belongs to G, the incrementary ratio is $\geqq 0$, we should find that

$$g(b) - g(a) > \tfrac{1}{2}(\iota - 1) k \{m(E) - \epsilon\} - p. \quad \text{............}(2)$$

The inequalities (1), (2) are in contradiction if

$$\tfrac{1}{2}km(E) > \epsilon\left(\iota k + \tfrac{k}{2} + 2N\right),$$

which will hold provided ϵ have been chosen sufficiently small. It follows that the set S cannot have measure > 0. The method of the above proof can be applied to shew that a similar result holds for $D^-f(x)$, $D_-f(x)$. It thus appears that, in the theorem (3), any pair of derivatives may be taken.

298. It has now been shewn that, in the set S_1, of points for which all four derivatives are finite, the upper and lower derivatives on the right are equal, and the upper and lower derivatives on the left are equal, except in a part of S_1, of measure zero; and thus, in a set of measure $m(S_1)$, there are unique derivatives both on the right and on the left. The points of S_1 at which these are unequal, form a set of measure zero, as is seen from the theorem that the points at which the upper derivative on one side is less than the lower derivative on the other side form a set of measure zero. We have accordingly the following theorem:

The set of points at which the four derivatives are all finite, and not all equal, has measure zero. Thus, in the set of points at which all the derivatives are finite, there is a finite differential coefficient everywhere, except at points of a set of measure zero.

Next, let us consider the set S_2, in which one derivative is infinite, $+\infty$, or $-\infty$, and the others are all finite. By theorem (1), S_2 has measure zero.

If, at points of a set S_3, two of the derivatives are infinite and the other two finite, then, except in the points of a part of the set, of which the measure is zero, these derivatives must be the upper derivative on one side and the lower derivative on the other side, and the two finite derivatives are equal.

If three of the derivatives are infinite at points of a set S_4, then, by the theorems (1) and (2), all four derivatives are infinite at all points of S_4, except in a part of the set, of which the measure is zero.

If all four derivatives are infinite at the points of a set S_4, then, except in a part of S_4, of which the measure is zero,

$$D^+f(x) = D^-f(x) = +\infty, \text{ and } D_+f(x) = D_-f(x) = -\infty.$$

The general result may be stated as follows:

If $f(x)$ be a measurable function, finite at each point, and if a set of points of measure zero be left out of account, then at every point x, either

(1). there is a finite differential coefficient, or

(2). the two upper derivatives are $+\infty$, and the two lower derivatives are $-\infty$, or

(3). the upper derivative on one side is $+\infty$, the lower derivative on the other side is $-\infty$, and the other two derivatives are finite, and equal.

This general theorem was established by Denjoy* for the case of a continuous function, and by G. C. Young† for the case of any measurable function. The investigation here given is based upon the work of the latter writer.

The following is a consequence of the general theorem:

There is no geometrical distinction of right and left with regard to the four extreme derivatives of a measurable finite function, except possibly at points of a set of measure zero.

A number of special theorems are particular cases of the general theorem. Thus, we have the following:

If a measurable function have finite derivatives except at points of a set of measure zero, it has a finite differential coefficient except at points of a set of measure zero.

This theorem was first given by Montel‡, for the case of a continuous function.

A monotone function $f(x)$ has a finite differential coefficient at every point which does not belong to a set of measure zero.

For, in the case of a monotone function, all four derivatives are ≥ 0, or all four are ≤ 0, and thus the cases (2) and (3) of the general theorem cannot arise.

This theorem was first established§, in connection with the theory of Integration, by Lebesgue, for the case of a continuous function. It was extended by W. H. Young to the case of a monotone function which is not necessarily continuous‖. A proof has also been given¶ by Rajchman and Saks.

---

* *Journal de Math.* (7), vol. I (1915), p. 105.

† *Proc. Lond. Math. Soc.* (2), vol. XV (1916), p. 360. See also Banach, *Comptes Rendus*, vol. CLXXIII (1921), p. 457, and Saks, *Fundamenta Mat.* vol. V (1924), p. 98.

‡ *Comptes Rendus*, vol. CLV (1912), p. 1478.

§ *Leçons sur l'intégration* (1904), p. 128.

‖ *Quarterly Journal*, vol. XLII (1911), p. 79.

¶ *Fundamenta Mat.* vol. IV (1923), p. 204.

Since every function of bounded variation is the difference of two monotone functions, we have the theorem that:

A function of bounded variation has a finite differential coefficient at each point which does not belong to a set of measure zero.

The following special case of the general theorem was first given by Lusin:

The points at which a continuous function has an infinite differential coefficient (with fixed sign) form a set of measure zero.

It has been shewn[*] by A. N. Singh that, *if a function is everywhere continuous (or if its points of discontinuity form a non-dense set) there exists an everywhere dense set of points at which the function has a progressive differential coefficient, finite or infinite.*

299. We have hitherto assumed that $f(x)$, whether bounded or unbounded, is finite, for each value of x. But if it be assumed that there is an exceptional set of points, of measure zero, at which $f(x)$ is infinite, the general theorem will still hold good, as is seen by examining the proofs of the various theorems. It is assumed that, in the definition of a derivative $Df(x)$, the incrementary ratio, for a pair of points at which the function is infinite, with the same sign, is ignored. At a point where $f(x) = +\infty$, we have

$$D^+f(x) = D_+f(x) = -\infty, \text{ and } D^-f(x) = D_-f(x) = +\infty;$$

and at a point where $f(x) = -\infty$, we have

$$D^+f(x) = D_+f(x) = +\infty, \text{ and } D^-f(x) = D_-f(x) = -\infty.$$

It thus appears that the set of points at which the upper derivative on one side is less than the lower derivative on the other side contains the set of points at which the function is infinite

An extension has been made by G. C. Young[†] to the case in which the function is infinite at points of a set of which the measure is > 0.

It has been shewn by Denjoy that a continuous function can be constructed, in which all three cases of the general theorem present themselves for one function, or in which one or more of the cases may be absent.

An investigation has been given[‡] by Burkill of the properties of the derivatives of functions of intervals.

EXAMPLES

The following examples have been given by Denjoy:

1. Let G be a perfect non-dense set of measure > 0; let $f(x) = 0$, at all points of G; and in each complementary interval (a_n, b_n) of G, let $f(x) = \frac{1}{b_n - a_n}\{(x - a_n)(b_n - x)\}^{\frac{1}{2}}$. At each point of G,

$$D^+f(x) = +\infty, \quad D_+f(x) = 0, \quad D^-f(x) = 0, \quad D_-f(x) = -\infty.$$

---

[*] *Bulletin Calcutta Math. Soc.* vol. XVI (1925–26), p. 101.

[†] *Loc. cit.* pp. 378–384.

[‡] *Proc. Lond. Math. Soc.* (2), vol. XXII (1923), p. 293.

In each complementary interval $f(x)$ has the maximum value $\frac{1}{2}$, and thus the function is not continuous at the points of G.

2. Let G be a perfect non-dense set, of measure zero, in the interval (a, b). Let u_n be the length of the contiguous interval (a_n, b_n); and suppose that $u_1, u_2, \dots u_n, \dots$ are in descending order of magnitude. Let $h_1, h_2, \dots h_n, \dots$ be positive numbers such that $\Sigma h_n u_n$ is convergent, and let $y_n(x) = \frac{2}{\pi} h_n u_n \sin^{-1}\left(\frac{x - a_n}{u_n}\right)^{\frac{1}{2}}$, for $a_n \leq x \leq b_n$. The function

$$f(x) = (a\,\Sigma x)\, h_n u_n + \omega y_m(x),$$

where ω is zero if x is a point of G, and is 1, if x is interior to a contiguous interval; and $(a\,\Sigma x)\, h_n u_n$ denotes the sum of those parts of $\Sigma h_n u_n$ that are in the interval (a, x). The function $f(x)$ is continuous; it has a finite differential coefficient at every point that does not belong to G, and it has an infinite differential coefficient at all points of G.

3. With the same notation as in Ex. 2, let $f(x) = 0$, in all points of G, and let

$$f(x) = u_n^2\, y\left(\frac{x - a_n}{u_n}\right),$$

in the interval (a_n, b_n): where $y(x) = (1 - x)^2 x^{\frac{3}{2}} \sin^2 \frac{1}{x}$. The function $f(x)$ is continuous; it has a finite and continuous differential coefficient at all points not belonging to G. In all end-points of contiguous intervals $D^- f(x) = +\infty$; and the other three derivatives are all zero at all points of G.

4. Let E be a perfect set, metrically dense in itself; and let

$$y(x) = 4\{x(1 - x)\}^{\frac{3}{2}} \sin^2 \frac{\pi}{8x(1 - x)}.$$

Let $f(x) = 0$, at all points of E; in u_n let $f(x) = \rho_n^{\frac{3}{2}}\, y\left(\frac{x - a_n}{b_n}\right)$, where ρ_n is chosen as follows:

From the interval (a, b) remove the intervals $u_1, u_2, \dots u_{n-1}$; then u_n lies in one of the remaining segments of (a, b); let ρ_n be the length of this segment. In all points of E,

$$D^- f(x) = +\infty; \quad D_- f(x) = -\infty; \quad D_+ f(x) = D^- f(x) = 0.$$

5. Defining E as in Ex. 4, let $f(x) = 0$, in all points of E, and let

$$y_n(x) = \{\rho_n x(1 - x)\}^{\frac{3}{2}} \sin \frac{k}{x(1 - x)}.$$

In all points of E,

$$D^+ f(x) = D^- f(x) = +\infty, \quad D_+ f(x) = D_- f(x) = -\infty.$$

300. The following general theorem is due* to W. H. Young.

The points at which one at least of the four derivatives of any given function is infinite form an ordinary inner limiting set, if it exists. The set of such points is accordingly of power c, *when it contains a component dense in itself; and otherwise it is enumerable, or finite.*

It follows that the set of such points is a set of the second category, in case it be everywhere dense.

The special case of the theorem which arises when the function is continuous, and the set of points is everywhere dense, was given by Brodén†.

---

* *Arkiv för Matematik, Astronomi och Fysik,* vol. I (1903).

† *Acta Univ. Lund.* vol. XXXIII (1897), p. 31.

Let x_0 be a point at which one of the four derivatives is infinite, it being immaterial whether the other derivatives are infinite or finite. A sequence $x_1, x_2, \dots x_n, \dots$ converging to x_0, and on one side of it, can be found, which has the property that, corresponding to an arbitrarily large positive number σ, an integer m_1 can be found, such that

$$\left|\frac{f(x_n)-f(x_0)}{x_n-x_0}\right| > \sigma, \text{ for } n > m_1;$$

further, m' can be chosen so great that

$$|x_n - x_0| < \frac{1}{\sigma}, \text{ for } n > m' \geq m_1.$$

Let the intervals (x_0, x_n) be prolonged on the side beyond x_n, each being increased by $\frac{1}{\sigma - 1}$ of its length; and the whole set of intervals so constructed for every point x of the set at which a derivative is infinite may be called I_σ.

Let $\sigma_1, \sigma_2, \sigma_3, \dots$ be a set of values of σ which increase without limit, then the corresponding sets of intervals $I_{\sigma_1}, I_{\sigma_2}, \dots$ define an inner limiting set of points, to which all the points x of the given set belong; and it will be shewn that no other points belong to this inner limiting set. If possible let ξ be a point of the inner limiting set which does not belong to the given set of points at which a derivative is infinite. There is at least one interval of each of the sets $I_{\sigma_1}, I_{\sigma_2}, \dots$, such that ξ is an interior point of it; let such intervals be $\delta_1, \delta_2, \dots$, and let $\xi_1, \xi_2, \xi_3, \dots$ be points of the given set interior to these intervals. Let $\bar{\xi}_1, \bar{\xi}_2, \bar{\xi}_3, \dots$ be the end-points of the intervals on the sides of those intervals which were not lengthened. We have

$$\delta_\iota = \left(1 - \frac{1}{\sigma_\iota - 1}\right) . |\xi_\iota - \bar{\xi}_\iota| < \frac{1}{\sigma_\iota - 1};$$

thus the points $\xi_1, \xi_2, \xi_3, \dots$, and also the points $\bar{\xi}_1, \bar{\xi}_2, \dots$, form a sequence of which ξ is the limit.

Since $\left|\frac{f(\xi_\iota) - f(\bar{\xi}_\iota)}{\xi_\iota - \bar{\xi}_\iota}\right| > \sigma_\iota$, therefore $|f(\xi_\iota) - f(\bar{\xi}_\iota)| > (\sigma_\iota - 1)\,\delta_\iota$.

Also a positive number A can be determined, such that for all values of ι,

$$\left|\frac{f(\xi) - f(\xi_\iota)}{\xi - \xi_\iota}\right| < A,$$

for otherwise ξ would be a point with an infinite derivative; and from this we see that

$$|f(\xi) - f(\xi_\iota)| < A\delta_\iota.$$

For a sufficiently great value of ι,

$$\sigma_\iota - 1 > A;$$

hence, for such a value of ι,

$$|f(\xi) - f(\bar{\xi}_\iota)| > (\sigma_\iota - A - 1)\,\delta_\iota$$

and thus

$$\left|\frac{f(\xi) - f(\bar{\xi}_\iota)}{\xi - \bar{\xi}_\iota}\right| > \sigma_\iota - A - 1.$$

Now $\sigma_\iota - A - 1$ is arbitrarily large for a sufficiently great ι; hence, since ξ is the limiting point of the sequence $\{\bar{\xi}_\iota\}$, there is an infinite derivative at ξ. This is contrary to the hypothesis made; therefore the points of the given set constitute the inner limiting set which has been defined.

FUNCTIONS OF TWO VARIABLES

301. It has already been shewn that many of the fundamental conceptions and theorems relating to the continuity, uniform continuity, &c., of functions of one variable hold good, without essential change, for functions of two or more variables.

Most of the points in which the theory of functions of a number of variables involves considerations which are not an immediate generalization of those which occur in the case of functions of a single variable are sufficiently illustrated by the case of functions of two variables. Accordingly the properties of functions of two variables will be considered in some detail.

That a function $f(x, y)$ should be continuous at a point (a, b), which is a limiting point of its domain, it is necessary, but not sufficient, that the function $f(x, b)$, of x, should be continuous at the point $x = a$, and that the function $f(a, y)$, of y, should be continuous at the point $y = b$. Thus a function may be continuous at a point with respect to x, and also with respect to y, whilst it is discontinuous with respect to the two-dimensional domain (x, y).

It is not even sufficient to ensure the continuity of $f(x, y)$ at a point, that it be continuous in every direction from the point. Thus

$$f(a + r\cos\theta,\ b + r\sin\theta)$$

may be a continuous function of r, at $r = 0$, for each value of θ in the interval $(0, 2\pi)$, and yet* the function may be discontinuous at (a, b).

The necessary and sufficient condition that $f(x, y)$ may be continuous at (a, b) may be expressed in the form that $f(x, y)$ *must be continuous in every direction at the point, and uniformly so for all directions.*

Thus, if $f(a + r\cos\theta,\ b + r\sin\theta)$ be continuous at $r = 0$, for each value of θ, and uniformly so for all values of θ, then, if ϵ be a prescribed positive number, a number ρ can be determined, independent of θ, such that

$$|f(a + r\cos\theta,\ b + r\sin\theta) - f(a, b)| < \epsilon,$$

---

* See Thomae, *Abriss einer Theorie der komplexen Functionen*, 2nd ed. p. 15.

provided $r < \rho$. From this condition it follows that $|f(x, y) - f(a, b)| < \epsilon$, provided $|x - a|$, $|y - b|$ are each $< \rho\sqrt{2}$, and thus the condition of continuity of the function is satisfied.

The remarks which have been made as regards the continuity of a function at a point are applicable without essential change if those functional values in the neighbourhood of the point are alone taken into account, which are in one of the four quadrants, the values at points on the axes bounding the quadrant being either included or excluded from consideration, as may be agreed upon. Thus the condition of continuity at a point may be satisfied for one such quadrant, and not for another one.

As regards the points of discontinuity of a function of two or more variables reference may be made to a memoir* by W. H. Young.

EXAMPLES

1. Let $f(x, y) = \frac{xy}{x^2 + y^2}$, and $f(0, 0) = 0$. This function is discontinuous at the point $(0, 0)$, although it is continuous at that point with respect to x, and also with respect to y, since $f(x, 0) = 0$, $f(y, 0) = 0$. In all other directions the function is discontinuous, for putting $x = r\cos\theta$, $y = r\sin\theta$, the function is $\frac{1}{2}\sin 2\theta$ and therefore has a constant value different from zero on a straight line for which θ is constant, unless θ has one of the values 0, $\frac{1}{2}\pi$, π, or $\frac{3}{2}\pi$.

2. Let† $f(x, y) = \frac{xy^2}{x^2 + y^4}$, $f(0, 0) = 0$. This function is discontinuous at the point $(0, 0)$, although it is continuous in each particular direction at that point. We find that $\frac{r\sin^2\theta}{\cos^2\theta + r^2\sin^4\theta} < \epsilon$, if $r < \frac{1}{2\epsilon}\operatorname{cosec}^2\theta\,\{1 - \sqrt{1 - 4\epsilon^2\cos^2\theta}\}$, and in order that this condition may be satisfied, the greatest value of r diminishes indefinitely as θ approaches the value $\frac{1}{2}\pi$; whereas when $\theta = \frac{1}{2}\pi$, the function is, for every value of r, equal to $f(0, 0)$. It is thus seen that the convergence in different directions is non-uniform.

DOUBLE AND REPEATED LIMITS

302. Let (a, b) be a limiting point of the domain for which a function $f(x, y)$ is defined, and let a neighbourhood, of which the corners are the four points $(a \pm \epsilon, b \pm \epsilon')$, be taken. Let U, L be the upper and lower boundaries of the function for all points of the domain in this neighbourhood, the functional value at (a, b) being however disregarded, in case (a, b) belongs to the domain. If ϵ, ϵ' be diminished, the number U cannot increase; and when values of ϵ, ϵ' belonging to sequences $\epsilon_1, \epsilon_2, \dots \epsilon_n, \dots$, and $\epsilon_1', \epsilon_2', \dots \epsilon_n', \dots$, each of which converges to the limit zero, are taken, and U_n be the value of U corresponding to the values ϵ_n, ϵ_n', of ϵ, ϵ', the numbers $U_1, U_2, \dots U_n, \dots$ form a sequence of numbers which do not increase. This sequence has then a limit $\bar{U}$, which may however have the

---

[*] W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. VIII (1909), p. 117.

[†] Genocchi-Peano, *Calc. Diff.* § 123.

improper value ∞, in case all the numbers U_n have this improper value. It is easily seen that $\overline{U}$ is independent of the particular sequences chosen for ϵ, ϵ'. This number $\overline{U}$ is said to be the *upper limit of the function at* (a, b), and may be denoted by

$$\overline{\lim_{x \sim a,\, y \sim b}} f(x, y).$$

The lower limit $\underline{\lim}_{x \sim a,\, y \sim b} f(x, y)$ may be defined in a similar manner, as the limit of a sequence of values of L; and it may have the improper value $-\infty$.

At a point of continuity of the function, the condition

$$\overline{\lim_{x \sim a,\, y \sim b}} f(x, y) = \underline{\lim}_{x \sim a,\, y \sim b} f(x, y)$$

is satisfied; and further, each of these limits is equal to $f(a, b)$, in case (a, b) belongs to the domain of the function.

Corresponding pairs of limits may also be defined for the case in which the functional values in one quadrant only are taken into account, the functional values on the axes being either included or excluded, in case they exist, as may be agreed upon.

The *saltus*, or *measure of discontinuity*, at the point (a, b) is measured by the excess of the greatest over the least of the three numbers

$$f(a, b),\quad \overline{\lim_{x \sim a,\, y \sim b}} f(x, y),\quad \underline{\lim}_{x \sim a,\, y \sim b} f(x, y).$$

The saltus at a point of discontinuity may have a finite value, or it may be indefinitely great.

In case $\overline{\lim}_{x \sim a,\, y \sim b} f(x, y) = \underline{\lim}_{x \sim a,\, y \sim b} f(x, y)$, their common value may be denoted by $\lim_{x \sim a,\, y \sim b} f(x, y)$, and the function is then said to have a definite *double limit* at the point (a, b); this double limit $\lim_{x \sim a,\, y \sim b} f(x, y)$ may be finite, or infinite with a definite sign.

When the upper and lower limits have different values, $\lim f(x, y)$ is frequently regarded as existent but indeterminate, the upper and lower limits being regarded as its limits of indeterminancy.

303. In considering the functional values in the neighbourhood of a point, and the functional limits at the point, it is frequently convenient to consider one quadrant only; this we may take, without loss of generality, to be the quadrant in which $x - a > 0$, $y - b > 0$. The results which will be established are essentially applicable to any one of the four quadrants, and can be immediately extended to the case in which account is taken of the whole neighbourhood of (a, b), by taking the totality of the results for the four separate quadrants, and for the lines $x = a$, $y = b$.

Assuming that $x - a > 0$, $y - b > 0$, the function $f(x, y)$ considered as a function of y only, with x constant, has two functional limits $\overline{f(x, b + 0)}$, $\underline{f(x, b + 0)}$, at the point (x, b); these may be denoted by $\overline{\lim}_{y\sim b} f(x, y)$, $\underline{\lim}_{y\sim b} f(x, y)$ respectively. In case these two limits are identical, their common value may be denoted by $\lim_{y\sim b} f(x, y)$, the functional limit $f(x, b + 0)$ having in that case a definite value.

If either of the limits $\overline{\lim}_{y\sim b} f(x, y)$, $\underline{\lim}_{y\sim b} f(x, y)$ is to be taken indifferently, we may denote them by $\overline{\underline{\lim}}_{y\sim b} f(x, y)$. This may be regarded as a function of x, such that its value at the point (x, b) is multiple-valued, and has $\overline{\lim}_{y\sim b} f(x, y)$, $\underline{\lim}_{y\sim b} f(x, y)$ for its limits of indeterminancy.

It may happen that $\overline{\underline{\lim}}_{y\sim b} f(x, y)$, considered as a function of x, has a definite functional limit at the point $x = a$; this limit may be either finite, or infinite with fixed sign. In case such a limit exists, it is denoted by $\lim_{x\sim a} \lim_{y\sim b} f(x, y)$, and is said to be the *repeated limit* of $f(x, y)$ at the point (a, b), the order of the limits being, that the limit for $y \sim b$ is taken first, and then afterwards the limit for $x \sim a$.

In case this repeated limit does not exist, either as a definite number, or as infinite with fixed sign, we may regard $\lim_{x\sim a} \lim_{y\sim b} f(x, y)$ as indeterminate, its limits of indeterminancy being

$$\overline{\lim}_{x\sim a}\, \overline{\lim}_{y\sim b} f(x, y), \text{ and } \underline{\lim}_{x\sim a}\, \underline{\lim}_{y\sim b} f(x, y).$$

The repeated limit $\lim_{y\sim b} \lim_{x\sim a} f(x, y)$, in which the limit with respect to x is first taken, and afterwards that with respect to y, may be defined in a precisely similar manner.

It is clear that the functional values on the straight lines $x = a$, $y = b$ are irrelevant as regards the existence, or the values, of the repeated limits.

In case the *double limit* $\lim_{x\sim a,\, y\sim b} f(x, y)$, for $x > a$, $y > b$, exists at the point (a, b), having either a finite value, or being infinite with fixed sign, the existence of the two repeated limits

$$\lim_{x\sim a} \lim_{y\sim b} f(x, y), \quad \lim_{y\sim b} \lim_{x\sim a} f(x, y)$$

follows as a consequence, their common value being $\lim_{x\sim a,\, y\sim b} f(x, y)$. In this case $\lim_{x\sim a} f\{x, \phi(x)\}$ also exists, and is equal to the double limit, where

$\phi(x)$ is any function of x, which is $> b$, and is such that $\lim\limits_{x \sim a} \phi(x) = b$. Also $\lim\limits_{t \sim \infty} f\{\phi(t), \psi(t)\}$ exists, and is equal to the double limit; where $\phi(t)$, $\psi(t)$ are functions of a variable t, such that $\phi(t) > a$, $\psi(t) > b$, and that $\lim\limits_{t \sim \infty} \phi(t) = a$, $\lim\limits_{t \sim \infty} \psi(t) = b$.

The converse of these statements does not hold good. In particular, the existence of $\lim\limits_{x \sim a,\, y \sim b} f(x, y)$ is not necessary either for the existence, with definite values, or for the equality, of the two repeated limits

$$\lim_{x \sim a} \lim_{y \sim b} f(x, y); \quad \lim_{y \sim b} \lim_{x \sim a} f(x, y).$$

EXAMPLES

1. Let $f(x, y)$ be defined for the positive quadrant by $f(x, y) = \dfrac{x - y}{x + y}$. We find

$$\lim_{x \sim 0} \lim_{y \sim 0} f(x, y) = 1, \quad \lim_{y \sim 0} \lim_{x \sim 0} f(x, y) = -1;$$

thus $\lim\limits_{x \sim 0,\, y \sim 0} f(x, y)$ cannot exist.

2. Let $f(x, y) = \dfrac{x^2 y^2}{x^2 y^2 - (x - y)^2}$. In this case $\lim\limits_{x \sim 0} \lim\limits_{y \sim 0} f(x, y)$ and $\lim\limits_{y \sim 0} \lim\limits_{x \sim 0} f(x, y)$ are both zero, and yet $\lim\limits_{x \sim 0,\, y \sim 0} f(x, y)$ does not exist; for if $y = x$, $f(x, y) = 1$; and therefore $\lim\limits_{x \sim 0} f(x, x) = 1$.

3. Let* $f(x, y)$ be defined for $x > 0$, $y > 0$, by the expression $(x + y) \sin\dfrac{1}{x} \sin\dfrac{1}{y}$. In this case $\overline{\lim\limits_{y \sim 0}}\, f(x, y) = x \sin\dfrac{1}{x}$, $\underline{\lim\limits_{y \sim 0}}\, f(x, y) = -x \sin\dfrac{1}{x}$, and $\overline{\lim\limits_{y \sim 0}}\, f(x, y) - \underline{\lim\limits_{y \sim 0}}\, f(x, y)$ has for $x = 0$ the limit zero. We have then $\lim\limits_{x \sim 0} \lim\limits_{y \sim 0} f(x, y) = 0$, since $x \sin\dfrac{1}{x}$, $-x \sin\dfrac{1}{x}$ have each the limit zero for $x = 0$. It is clear that $\lim\limits_{y \sim 0} \lim\limits_{x \sim 0} f(x, y)$ is also zero. If $0 < x < \frac{1}{2}\epsilon$, and $0 < y < \frac{1}{2}\epsilon$, we see that $|f(x, y)| < \epsilon$, and therefore $\lim\limits_{x \sim 0,\, y \sim 0} f(x, y)$ exists, and is equal to zero.

304. An important matter for investigation is the determination of the necessary and sufficient conditions for the existence and equality of the two repeated limits at a point. A knowledge of such conditions, as also of sufficient conditions, is required in various fundamental theorems of analysis which turn upon the legitimacy of inverting the order of a repeated limiting process.

It will be observed that the existence of $\lim\limits_{x \sim a} \lim\limits_{y \sim b} f(x, y)$ does not necessarily involve the existence of $\lim\limits_{y \sim b} f(x, y)$ as a definite number, since $\lim\limits_{x \sim a} \overline{\lim\limits_{y \sim b}}\, f(x, y)$, $\lim\limits_{x \sim a} \underline{\lim\limits_{y \sim b}}\, f(x, y)$ may both exist and have the same value, without it being necessarily the case that $\overline{\lim\limits_{y \sim b}}\, f(x, y)$, $\underline{\lim\limits_{y \sim b}}\, f(x, y)$ are identical. It is however necessary that $\overline{\lim}_{y\sim b} f(x, y) - \underline{\lim}_{y\sim b} f(x, y)$ should converge to the limit zero, as x converges to the value a.

---

* Pringsheim, *Encyclopädie der math. Wissensch.* II A, 1, p. 51.

The necessary and sufficient conditions required are contained in the following general theorem*.

In order that the repeated limits $\lim_{x\sim a} \lim_{y\sim b} f(x, y)$, $\lim_{y\sim b} \lim_{x\sim a} f(x, y)$ *may both exist, and have the same finite value, it is necessary and sufficient,* (1), *that* $\overline{\lim}_{y\sim b} f(x, y) - \underline{\lim}_{y\sim b} f(x, y)$ *should have the limit zero, for* $x \sim a$, *and that* $\overline{\lim}_{x\sim a} f(x, y) - \underline{\lim}_{x\sim a} f(x, y)$ *should have the limit zero, for* $y \sim b$; *and* (2), *that, corresponding to any fixed positive number* ϵ *arbitrarily chosen, a positive number* β *can be determined, satisfying the condition that, for each value of* y *interior to the interval* $(b, b + \beta)$, *a positive number* α_y, *in general dependent on* y, *exists, such that, for this value of* y, $f(x, y)$ *lies between* $\overline{\lim}_{y\sim b} f(x, y) + \epsilon$ *and* $\underline{\lim}_{y\sim b} f(x, y) - \epsilon$, *for all values of* x *interior to the interval* $(a, a + \alpha_y)$.

Let us first assume that the conditions stated in the theorem are satisfied. A value of y may, in virtue of (1), be so chosen that the difference of the two limits $\overline{\lim}_{x\sim a} f(x, y)$, $\underline{\lim}_{x\sim a} f(x, y)$ is less than an arbitrarily chosen number η; and this value of y may also be so chosen that it is interior to $(b, b + \beta)$. For this fixed value of y, an interval $(a, a + \alpha_y')$, for x, may be so chosen that $f(x, y)$ lies between $\overline{\lim}_{x\sim a} f(x, y) + \epsilon$ and $\underline{\lim}_{x\sim a} f(x, y) - \epsilon$, provided y has the fixed value, and $a < x < a + \alpha_y'$; this follows from the definition of the upper and lower limits. Again, from the condition (1), a number α'' can be determined, such that, if x be interior to the interval $(a, a + \alpha'')$, the difference between the two limits $\overline{\lim}_{y\sim b} f(x, y)$, $\underline{\lim}_{y\sim b} f(x, y)$ is less than η. Now let $\overline{\alpha_y}$ be the smallest of the three numbers α_y, α_y', α''; then, if x_1, x_2 be any two values of x within the interval $(a, a + \overline{\alpha_y})$, and y have the fixed value; by applying the conditions of the theorem, we see that the conditions

$$|f(x_1, y) - f(x_2, y)| < \eta + 2\epsilon,$$

$$\left| f(x_1, y) - \overline{\lim_{y\sim b}} f(x_1, y) \right| < \eta + \epsilon,$$

$$\left| f(x_2, y) - \overline{\lim_{y\sim b}} f(x_2, y) \right| < \eta + \epsilon,$$

---

* This theorem was given by Hobson, *Proc. Lond. Math. Soc.* (2), vol. v (1907), p. 226. It was transformed into an equivalent statement by Martinotti, *Rendiconti del circ. math. di Palermo*, vol. xxxvii (1914), p. 19. On the whole subject see also the Dissertation by J. Ridder, *Enkele algemeene limietstellingen met toepassingen op reële functies*, Utrecht (1921).

are all satisfied. It follows that

$$\left| \overline{\lim_{y\sim b}} f(x_1, y) - \overline{\lim_{y\sim b}} f(x_2, y) \right| < 3\eta + 4\epsilon$$

for every pair of points x_1, x_2 within the interval $(a, a + \overline{a}_y)$. Hence, since ϵ, η are both arbitrarily small, $\overline{\lim\limits_{y\sim b}} f(x, y)$ converges, for $x \sim a$, to a definite value which is the limit both of $\overline{\lim\limits_{y\sim b}} f(x, y)$ and of $\underline{\lim\limits_{y\sim b}} f(x, y)$, when $x \sim a$; and thus $\lim\limits_{x\sim a} \lim\limits_{y\sim b} f(x, y)$ exists.

Again, since $\lim\limits_{x\sim a} \lim\limits_{y\sim b} f(x, y)$ has a definite value, an interval $(a, a + \delta)$ can be determined, such that, for any point x interior to it,

$$\left| \lim_{x\sim a} \lim_{y\sim b} f(x, y) - \overline{\lim_{y\sim b}} f(x, y) \right| < \epsilon.$$

Now $\lim\limits_{x\sim a} \lim\limits_{y\sim b} f(x, y) - \overline{\lim\limits_{x\sim a}} f(x, y)$ is the sum of the three differences

$$\lim_{x\sim a} \lim_{y\sim b} f(x, y) - \overline{\lim_{y\sim b}} f(x, y),$$

$$\overline{\lim_{y\sim b}} f(x, y) - f(x, y), \quad f(x, y) - \overline{\lim_{x\sim a}} f(x, y),$$

and for a fixed y, chosen as before, x may be chosen so that it not only lies within the interval $(a, a + \delta)$, but is also such that

$$\left| f(x, y) - \overline{\lim_{y\sim b}} f(x, y) \right|, \quad \left| f(x, y) - \overline{\lim_{x\sim a}} f(x, y) \right|$$

are each less than $\eta + 2\epsilon$. It follows that

$$\left| \lim_{x\sim a} \lim_{y\sim b} f(x, y) - \overline{\lim_{x\sim a}} f(x, y) \right| < 5\epsilon + 2\eta,$$

and thus that $\overline{\lim\limits_{x\sim a}} f(x, y)$ converges, as y converges to b, to the limit $\lim\limits_{x\sim a} \lim\limits_{y\sim b} f(x, y)$. It has thus been shewn that the two repeated limits both exist, and have the same value.

Conversely, let us assume that the repeated limits both exist, and are finite and equal. We have then $\left| \lim\limits_{x\sim a} \lim\limits_{y\sim b} f(x, y) - \overline{\lim\limits_{x\sim a}} f(x, y) \right| < \zeta$, provided y lies between b and $b + \beta$, where β is some fixed number, ζ being an arbitrarily chosen positive number; from this it follows that

$$\overline{\lim_{x\sim a}} f(x, y) - \underline{\lim_{x\sim a}} f(x, y)$$

is $< 2\zeta$, for $b < y < b + \beta$. Also

$$\left| \overline{\lim_{y\sim b}} f(x, y) - \lim_{x\sim a} \lim_{y\sim b} f(x, y) \right| < \zeta,$$

provided x lies within some fixed interval $(a, a + \delta')$; and from this it follows that $\overline{\lim}_{y\sim b} f(x, y) - \underline{\lim}_{y\sim b} f(x, y)$ is $< 2\zeta$, for $a < x < a + \delta'$. Since ζ is arbitrarily small, we now see that the condition (1) of the theorem is satisfied. Further we see that

$$\left| f(x, y) - \overline{\lim}_{x\sim a} f(x, y) \right| < 2\zeta - \zeta',$$

where ζ' is any arbitrarily chosen positive number, provided x lies within some interval $(a, a + \alpha_y')$, where α_y' depends upon y, and may diminish indefinitely as y approaches the value b. It follows from the three inequalities, that

$$\left| f(x, y) - \overline{\lim}_{y\sim b} f(x, y) \right| < 4\zeta - \zeta',$$

provided $b < y < b + \beta$, and provided also x lies within some interval $(a, a + \alpha_y)$, where α_y depends in general upon y. Since ζ and ζ' are both arbitrarily small, it follows that the condition (2) of the theorem is satisfied.

If the condition (2), in the above general theorem, be replaced by the more stringent condition that, corresponding to any fixed positive number ϵ, arbitrarily chosen, a positive number β can be determined, which is such that, for each value of y interior to the interval $(b, b + \beta)$, a positive number α_y, dependent on y, exists, such that, for this value of y, *and for all smaller values*, $f(x, y)$ lies between $\overline{\lim}_{y\sim b} f(x, y) - \epsilon$ and $\underline{\lim}_{y\sim b} f(x, y) - \epsilon$, then this condition and the condition (1) are the necessary and sufficient conditions that not only $\lim_{x\sim a}\lim_{y\sim b} f(x, y)$, $\lim_{y\sim b}\lim_{x\sim a} f(x, y)$ exist and are equal, but also that the double limit $\lim_{x\sim a,\, y\sim b} f(x, y)$ exists, having a definite value, the same as the repeated limits. In case the function be defined for values of x, y on the lines $x = a$, $y = b$, the additional conditions must be added that the functional values on these lines also converge to the same limit $\lim_{x\sim a,\, y\sim b} f(x, y)$.

For, under the conditions stated, we have, provided y lies within the interval $(b, b + \beta_1)$, where $\beta_1 < \beta$,

$$\left| f(x, y) - \overline{\lim}_{y\sim b} f(x, y) \right| < \epsilon - \eta,$$

where x has any value in the interval $(a, a + \zeta)$, ζ being the lesser of the two numbers α_{β_1} and δ'; the number δ' being so chosen that

$$\left| \overline{\lim}_{y\sim b} f(x, y) - \underline{\lim}_{y\sim b} f(x, y) \right| < \eta, \text{ for } a < x < a + \delta'.$$

Also $\left| \overline{\lim}_{y\sim b} f(x, y) - \lim_{x\sim a}\lim_{y\sim b} f(x, y) \right| < \epsilon$, provided x lies within an interval chosen sufficiently small. Hence the condition

$$\left| f(x, y) - \lim_{x\sim a}\lim_{y\sim b} f(x, y) \right| < 2\epsilon + \eta$$

is satisfied, provided $b < y < b + \beta_1$, and provided x lies within an interval of which the length may depend upon ϵ and η. It follows, since ϵ, η are arbitrarily small, that $f(x, y)$ has a definite double limit at the point (a, b). That the conditions stated are necessary, follows at once from the definition of $\lim\limits_{x\sim a,\, y\sim b} f(x, y)$.

305. The theorem obtained in § 304 may be simplified in the case in which $\lim\limits_{y\sim b} f(x, y)$, $\lim\limits_{x\sim a} f(x, y)$ both have definite values at all points on the straight lines $x = a$, $y = b$ which are in sufficiently small neighbourhoods of the point (a, b). We may then state the theorem as follows:

If $\lim\limits_{y\sim b} f(x, y)$, $\lim\limits_{x\sim a} f(x, y)$ *have definite finite values in the neighbourhood of the point* (a, b), *then the necessary and sufficient condition that the two repeated limits* $\lim\limits_{x\sim a}\lim\limits_{y\sim b} f(x, y)$, $\lim\limits_{y\sim b}\lim\limits_{x\sim a} f(x, y)$ *may both exist, and have the same finite value, is that, corresponding to any fixed positive number* ϵ, *arbitrarily chosen, a positive number* β *can be determined, which is such that, for each value of* y *interior to the interval* $(b, b + \beta)$, *a positive number* α_y, *in general dependent on* y, *exists, such that, for this value of* y,

$$\left| f(x, y) - \lim_{y\sim b} f(x, y) \right| < \epsilon,$$

for all values of x *within the interval* $(a, a + \alpha_y)$.

In case the condition $\left| f(x, y) - \lim\limits_{y\sim b} f(x, y) \right| < \epsilon$, for all values of x within $(a, a + \alpha_y)$, be satisfied, not only for the particular value of y, but for all smaller values, and this hold for every ϵ, then the double limit $\lim\limits_{x\sim a,\, y\sim b} f(x, y)$ exists, and is equal to each of the repeated limits. In this case the point (a, b) is said to be *a point of uniform convergence* of the function $f(x, y)$ to the limit $\lim\limits_{y\sim b} f(x, y)$, with respect to the parameter x; and thus, for such a point, there exists, for each value of ϵ, an interval $(a, a + \alpha)$, where α depends in general upon ϵ, such that, for each value of x within this interval, the condition $\left| f(x, y) - \lim\limits_{y\sim b} f(x, y) \right| < \epsilon$ is satisfied, provided y be less than some fixed value which is the same for the whole x-interval $(a, a + \alpha)$.

It may happen that, as ϵ is indefinitely diminished, α has a positive minimum $\bar{\alpha}$. In that case the fixed interval $(a, a + \bar{\alpha})$ is such that, for each ϵ, the condition $\left| f(x, y) - \lim\limits_{y\sim b} f(x, y) \right| < \epsilon$ is satisfied for all values of x within the fixed interval $(a, a + \bar{\alpha})$, provided y is less than some fixed value dependent on ϵ, the same for the whole x-interval. In this case $f(x, y)$ is said to *converge to* $\lim\limits_{y\sim b} f(x, y)$ *uniformly within the interval* $(a, a + \bar{\alpha})$, with respect to the parameter x. Not only the point (a, b), but also each interior point of the interval $(a, a + \alpha)$, is then a point of uniform convergence of $f(x, y)$ to $\lim\limits_{y \sim b} f(x, y)$, with respect to the parameter x.

306. The necessary and sufficient conditions for the existence and equality of the two repeated limits of $f(x, y)$, at (a, b), may be put into the following form, different from that of the theorem of § 304:

The necessary and sufficient conditions that

$$\lim_{x \sim a} \lim_{y \sim b} f(x, y) = \lim_{y \sim b} \lim_{x \sim a} f(x, y),$$

their value being finite, are (1), *that* $\overline{\lim\limits_{x \sim a}} f(x, y)$ *converge to a definite value* $\lim\limits_{y \sim b} \lim\limits_{x \sim a} f(x, y)$, *when* y *converges to* b, *and that* $\overline{\lim\limits_{x \sim a}} f(x, y) - \underline{\lim\limits_{x \sim a}} f(x, y)$ *converge to zero, for* $x \sim a$; *and* (2), *that, corresponding to any arbitrarily chosen positive number* ϵ, *and to an arbitrarily chosen value* β_0 *of* β, *a value* $y_1 < b + \beta_0$, *of* y, *can be found, and also a positive number* α, *such that the condition that* $f(x, y_1)$ *lies between*

$$\overline{\lim_{y \sim b}} f(x, y) + \epsilon, \text{ and } \underline{\lim_{y \sim b}} f(x, y) - \epsilon$$

is satisfied for every value of x *within the interval* $(a, a + \alpha)$.

In case $\lim\limits_{y \sim b} f(x, y)$ *everywhere exists in the neighbourhood of* (a, b), *the condition* (2) *is that* $\left| f(x, y_1) - \lim\limits_{y \sim b} f(x, y) \right| < \epsilon$, *for every value of* x *within the interval* $(a, a + \alpha)$.

That the conditions contained in the theorem are necessary is seen from the theorem of § 304; it will be shewn that they are sufficient. Let us assume that the conditions are satisfied. We have

$$\overline{\lim_{y \sim b}} f(x, y) - \lim_{y \sim b} \lim_{x \sim a} f(x, y) = \left[\overline{\lim_{y \sim b}} f(x, y) - f(x, y)\right] + \left[f(x, y) - \overline{\lim_{x \sim a}} f(x, y)\right] + \left[\overline{\lim_{x \sim a}} f(x, y) - \lim_{y \sim b} \lim_{x \sim a} f(x, y)\right].$$

A positive number β_1 can now be chosen, such that, if $b < y < b + \beta_1$, the condition $\left| \overline{\lim\limits_{x \sim a}} f(x, y) - \lim\limits_{y \sim b} \lim\limits_{x \sim a} f(x, y) \right| < \epsilon$ is satisfied, moreover we may choose β_1 so that it is $< \beta_0$.

Next, a value y_1, of y, exists, such that $f(x, y_1)$ lies between

$$\overline{\lim_{y \sim b}} f(x, y) + \epsilon, \text{ and } \underline{\lim_{y \sim b}} f(x, y) - \epsilon,$$

provided x be within the interval $(a, a + \alpha)$: the value of β_0 may be chosen so small that $\overline{\lim\limits_{x \sim a}} f(x, y) - \underline{\lim\limits_{x \sim a}} f(x, y) < \epsilon$, for every value of y which is

$< b - \beta_0$, and therefore for the value y_1, of y. Again, an interval for x, possibly less than $(a, a + \alpha)$, can be so chosen that

$$\overline{\lim_{y\sim b}}\, f(x, y) - \underline{\lim_{y\sim b}}\, f(x, y) < \epsilon,$$

provided x lies within the interval. It follows that an interval $(a, a + \alpha')$, for x, can be found, such that $\left|\overline{\underline{\lim}}_{y\sim b}\, f(x, y) - f(x, y_1)\right| < 3\epsilon$. Further, the interval within which x lies may, if necessary, be so restricted that

$$\left|f(x, y_1) - \overline{\underline{\lim}}_{x\sim a}\, f(x, y_1)\right| < 2\epsilon.$$

Hence, provided x lies within a definite interval, we see that

$$\left|\overline{\underline{\lim}}_{y\sim b}\, f(x, y) - \lim_{y\sim b} \lim_{x\sim a} f(x, y)\right| < 6\epsilon;$$

and since this condition holds for an arbitrary ϵ, it follows that $\overline{\underline{\lim}}_{y\sim b}\, f(x, y)$ converges, for $x \sim a$, to $\lim\limits_{y\sim b} \lim\limits_{x\sim a} f(x, y)$; and thus the sufficiency of the conditions is established.

THE LIMITS OF MONOTONE AND QUASI-MONOTONE FUNCTIONS OF TWO VARIABLES

307. A monotone function $f(x, y)$, of two variables x and y, was defined in § 253 as a function such that $f(x', y') \geqq f(x, y)$, or such that $f(x', y') \leqq f(x, y)$, for every pair of points (x, y), (x', y'), such that $x' \geqq x$, $y' \geqq y$. The following theorem, as regards the limits, at a point, of such a function, will be established.

If a function $f(x, y)$ is bounded and monotone in some neighbourhood of the point (a, b), then, in each of the two open quadrants $x > a$, $y > b$; $x < a$, $y < b$, the double limit of $f(x, y)$, at the point (a, b), has a definite value.

It will be observed that these limits are independent of the values of $f(x, y)$ on the straight lines through (a, b) parallel to the axes.

It follows from the theorem that each of the repeated limits

$$\lim_{x\sim a} \lim_{y\sim b} f(x, y), \quad \lim_{y\sim b} \lim_{x\sim a} f(x, y)$$

in one of the two open quadrants exists, and that they both have the value of the corresponding double limit.

We consider the values of the function in a cell Δ, of which (a, b) is an interior point, and such that $f(x, y)$ is monotone in the cell, and we take that part of the cell which is in the first quadrant $x > a$, $y > b$. Let $\phi_1(x, y)$ denote $\lim\limits_{k\sim 0} f(x, y + k)$, $k > 0$, and let $\phi_2(x, y)$ denote $\lim\limits_{h\sim 0} f(x, y - k)$; the functions $\phi_1(x, b)$, $\phi_2(x, b)$ are monotone functions of x.

We may suppose that $f(x', y') \geq f(x, y)$, when $x' \geq x$, $y' \geq y$, the other case may be treated in the same manner. We have then

$$\lim_{h\sim 0}\lim_{k\sim 0} f(a-h, b-k) = \lim_{h\sim 0}\phi_1(a-h, b) = \phi_1(a-0, b).$$

A value h_1, of h, can be so determined that $\phi_1(a-h_1, b) - \phi_1(a-0, b)$ is $< \frac{1}{2}\epsilon$, where ϵ is an arbitrarily chosen positive number: a value k_1, of k, can then be so determined that $f(a-h_1, b-k_1) - \phi_1(a-h_1, b) < \frac{1}{2}\epsilon$. We have then $0 \leq f(a-h, b-k) - \phi_1(a-0, b) < \epsilon$, for all values of h and k such that $0 < h \leq h_1$, $0 < k \leq k_1$. It follows, since ϵ is arbitrarily small, that the double limit of $f(a-h, b-k)$, as $h \sim 0$, $k \sim 0$ exists, and has the value $\phi_1(a-0, b)$. A similar proof can be given for the case of the other qudrant $x < a$, $y < b$.

The double limits in the other two quadrants $x > a$, $y < b$; $x < a$, $y > b$ do not necessarily exist. A theorem has however been given* by R. C. Young, from which, in the case of quasi-monotone functions, it follows that:

If a function $f(x, y)$ is bounded and quasi-monotone in some neighbourhood of the point (a, b), then, in each of the four open quadrants, $x > a$, $y > b$; $x < a$, $y < b$; $x > a$, $y < b$; $x < a$, $y > b$, the double limit of $f(x, y)$ at the point (a, b) has a definite value.

It is sufficient to consider the case of the quadrant $x > a$, $y > b$, when $\Delta_{(x, y)}^{(x', y')} f(x, y) \geq 0$, for any two points (x, y), (x', y'), in the cell

$$(a, b; a+h, b+k),$$

such that $x' > x$, $y' > y$, and the function $f(x, y)$ is of any of the four types specified in § 255. The other cases given there can all be treated in a precisely similar manner.

If P, Q denote the points (x, y), (x', y') respectively, we can denote $\Delta_{(x, y)}^{(x', y')} f(x, y)$ by $\Delta_P^Q f$, the points (a, b), $(a+h, b+k)$ may be denoted by A, B.

For any cell (P, Q) contained in (A, B), we have $\Delta_P^Q f \leq \Delta_A^B f$: for the cell (P, Q) may be one of a number of cells into which (A, B) is divided, and $\Delta_A^B f$ is equal to the sum, for all these cells (α, β), of $\Delta_\alpha^\beta f$: and $\Delta_\alpha^\beta f$ is by hypothesis ≥ 0.

If (A, Q_1), (A, Q_2), ... (A, Q_n), ... be a sequence of cells, such that Q_n is in the cell (A, Q_{n-1}), for every value of n, the sequence $\{\Delta_A^{Q_n} f\}$ is monotone non-increasing, and therefore converges to a definite limit, as $n \sim \infty$. Let $\{P_n\}$, $\{Q_n\}$ be any two sequences of points, each converging to A; a sequence $P_{n_1}, Q_{m_1}, P_{n_2}, Q_{m_2}, \ldots P_{n_\iota}, Q_{m_\iota}, \ldots$ can be so determined that Q_{m_ι} is in the cell (A, P_{n_ι}), and P_{n_ι} is in the cell $(A, Q_{m_{\iota-1}})$, for all values of ϵ: then the sequence $\Delta_A^{P_{n_1}} f, \Delta_A^{Q_{m_1}} f, \Delta_A^{P_{n_2}} f, \Delta_A^{Q_{m_2}} f, \dots$ has a unique limit. If $\{\Delta_A^{P_n} f\}, \{\Delta_A^{Q_m} f\}$ both have definite limits, it follows that these limits have the same value. Hence it is seen that $\Delta_A^P f$ has a definite limit as P converges to A through the points of any sequence of which A is the unique limiting point, and that the value of the limit is the same for all such sequences. That $\Delta_A^P f$ has a definite limit, in the sense that there exists a neighbourhood of A such that in that neighbourhood $\Delta_A^P f$ differs from its limit by less than an arbitrarily prescribed positive number ϵ, is proved as in § 211, in the discussion of the equivalence of the definitions of continuity given by Heine and by Cauchy, by a method which requires the use of the multiplicative axiom. Since

---

* *L'enseignement math.* 1924–5, p. 79.

$$f(x, y) + f(a, b) - f(x, b) - f(a, y)$$

converges to a definite limit as $x \sim a$, $y \sim b$; where (x, y) is in the open positive quadrant, and since $f(x, b), f(a, y)$ are both monotone functions, either of which may be non-increasing, or non-diminishing, according to the particular type of the quasi-monotone function, it follows, since $f(x, b), f(a, y)$, have in any case. definite limits as $x \sim a$, $y \sim b$, that $f(x, y)$ must have a definite double limit. Exactly similar proofs apply to each of the four quadrants, whatever be the type of the quasi-monotone function.

308. We now consider further the functions which have been defined in § 255, and named quasi-monotone functions. If $f(x, y)$ be such a function, the function $F(x, y)$ defined by

$$F(x, y) \equiv f(x, y) - f(x, b) - f(a, y) + f(a, b)$$

is a monotone function.

For a quasi-monotone function, the following theorem* relating to the points of discontinuity of the function, will be established:

If $f(x, y)$ be a quasi-monotone function, there exist two enumerable sets of straight lines parallel to the x-axis and the y-axis, respectively, such that every point of discontinuity of the function, with respect to (x, y), lies on a straight line belonging to one or other of the enumerable sets.

Let $\phi_1(x, y), \phi_2(x, y)$ be defined as in § 307; then, if $f(x, y)$ be monotone non-diminishing with respect to x, for each constant value of y, and also monotone non-diminishing with respect to y, for each constant value of x, we have

$$\phi_1(x - 0, y) \geqq f(x + 0, y) \geqq \phi_2(x + 0, y) \geqq f(x, y - 0) \geqq \phi_2(x - 0, y),$$
$$\phi_1(x + 0, y) \geqq f(x, y + 0) \geqq \phi_1(x - 0, y) \geqq f(x - 0, y) \geqq \phi_2(x - 0, y).$$

For a constant value α, of x, the discontinuities of $f(x, y)$ with respect to y are at points of an enumerable set on the straight line $x = \alpha$. Let b be the value of x on the right-hand boundary of the cell in which the function is defined. If $x = \alpha$, $y = \beta$ be a point at which $f(\alpha, y)$ is continuous with respect to y, then $f(x, y)$ is continuous with respect to y at $y = \beta$, for each value of x in the cell. For, let $\beta - k$, $\beta + k'$ be a neighbourhood of β, such that $0 \leq f(\alpha, \beta - k) - f(\alpha, \beta + k') < \epsilon$, then from the definition of $f(x, y)$, we see that $0 \leq f(x, \beta - k) - f(x, \beta + k') < \epsilon$. It then easily follows that $f(x, y)$ is continuous with respect to y, for $y = \beta$, for any fixed value of x. For such a value of y, as β, we have

---

* See W. H. Young. *Proc. Royal Soc.* vol. xciii (1917), p. 31, where these functions are termed "monotone functions."

$$\phi_1(x, y) = \phi_2(x, y),$$

for every value of x. This equality therefore holds for all values of y which do not belong to a certain enumerable set. Again $f(x, y)$ is a monotone non-diminishing function of x, for each value of y, and, as before, we see that it is a continuous function of x, for all values of x not belonging to a certain enumerable set, for each value of y.

If (x, y) is any point not on either set of straight lines parallel to the axis, defined by the exceptional values of x and of y, indicated above, we have $\phi_1(x + 0, y) = \phi_2(x - 0, y)$; and then, employing the inequalities given above, we see that the four limits, at (x, y), of the function in the open quadrants, and the four limits at (x, y) dependent on the values of the function on the parallels to the axes, through the point, are all equal; and therefore (x, y) is a point of continuity of the function.

In the case in which $f(x, y)$ is a non-diminishing function of y, for each constant value of x, and in which $f(x, y)$ is a non-increasing function of x, for each value of y; the inequalities employed above can be replaced by

$$\phi_2(x + 0, y) \leq f(x, y - 0) \leq \phi_1(x - 0, y) \leq f(x - 0, y) \leq \phi_2(x - 0, y),$$

$$\phi_2(x + 0, y) \leq f(x + 0, y) \leq \phi_1(x - 0, y) \leq f(x, y - 0) \leq \phi_1(x - 0, y);$$

and a similar argument to that above will shew that

$$\phi_2(x + 0, y) = \phi_1(x - 0, y)$$

at every point (x, y) not on either of two enumerable sets of lines parallel to the axes. The cases of functions of the other types can be treated in a similar manner.

PARTIAL DIFFERENTIAL COEFFICIENTS

309. If, at a point (x_0, y_0), in the domain for which the function $f(x, y)$ is defined, the limit $\lim\limits_{h \sim 0} \dfrac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$ exists, having either a definite finite value, or being indefinitely great, but of fixed sign, this limit is said to be the *partial differential coefficient* of $f(x, y)$, at (x_0, y_0), with respect to x; it is usually denoted by $\dfrac{\partial f(x_0, y_0)}{\partial x_0}$.

When the limit $\lim\limits_{k\sim 0} \dfrac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}$ exists, it is said to be the partial differential coefficient of $f(x, y)$, at (x_0, y_0), with respect to y; it is denoted by $\dfrac{\partial f(x_0, y_0)}{\partial y_0}$.

In general, h, k, in these definitions, are regarded as having either sign. It is possible that either of the above limits may not exist, but that there may be two definite limits, one for positive values of the increment h or k, and the other for negative values. In that case the two limits are said to be the progressive and regressive partial differential coefficients with respect to the particular variable. It is of course possible that, at a particular point, one of these may exist, and not the other.

That the two partial differential coefficients $\dfrac{\partial f}{\partial x_0}$, $\dfrac{\partial f}{\partial y_0}$ may exist, it is necessary, but not sufficient, that $f(x, y)$ should, at the point (x_0, y_0), be continuous with respect to x, and also with respect to y.

To express the increment $f(x_0 + h, y_0 + k) - f(x_0, y_0)$, of the function $f(x, y)$, when the two numbers x_0, y_0 receive increments h, k respectively, we have

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = [f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)] \\ + [f(x_0, y_0 + k) - f(x_0, y_0)].$$

If we now assume that $\dfrac{\partial f}{\partial y}$ exists at the point (x_0, y_0), and has a finite value, we have

$$\frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k} = \frac{\partial f(x_0, y_0)}{\partial y_0} + \sigma(k),$$

where $\sigma(k)$ converges to the limit zero, when k is indefinitely diminished.

Again, $\dfrac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h}$ converges to the limit $\dfrac{\partial f}{\partial x_0}$, when k is first diminished to the limit zero, and afterwards h converges to zero, it being assumed that $\dfrac{\partial f}{\partial x_0}$ has a definite value, and also that $f(x, y)$ is continuous with respect to y, for the value $y = y_0$, where x has any value in a neighbourhood of x_0. In order, however, that the double limit

$$\lim_{h\sim 0,\, k\sim 0} \frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h}$$

may exist, in which case its value is $\dfrac{\partial f}{\partial x_0}$, being independent of the mode in which h, k approach their limits, it is necessary and sufficient that

$$\frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h}$$

should be a continuous function of (h, k) at the point $h = 0$, $k = 0$.

If this condition be satisfied, positive numbers h_1, k_1 can be determined, such that

$$\left|\frac{f(x_0+h, y_0+k) - f(x_0, y_0+k)}{h} - \frac{\partial f}{\partial x_0}\right| < \eta,$$

where η is a prescribed positive number, and $0 < |h| < h_1$, $0 \leq |k| < k_1$.

We have now

$$\frac{f(x_0+h, y_0+k) - f(x_0, y_0+k)}{h} = \frac{\partial f}{\partial x_0} + \rho(h, k),$$

where $\rho(h, k)$ converges to zero, independently of the mode in which h, k converge to zero.

Under the conditions stated, we have

$$f(x_0+h, y_0+k) - f(x_0, y_0) = h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} + h\rho + k\sigma,$$

where ρ, σ converge to zero, when h and k are indefinitely diminished, independently of the mode in which they approach their limits. This is equivalent to the statement that, corresponding to an arbitrarily assigned positive number η, positive numbers h_1, k_1 can be determined so that $|\rho|$ and $|\sigma|$ are each $< \eta$, for all values of h and k such that $|h| < h_1$, $|k| < k_1$.

In the notation of differentials, denoting $f(x, y)$ by z, we have

$$dz_0 = \frac{\partial z}{\partial x_0}dx + \frac{\partial z}{\partial y_0}dy,$$

the expression on the right-hand side being termed the *total differential* of z at the point (x, y), and $\frac{\partial z}{\partial x_0}dx$, $\frac{\partial z}{\partial y_0}dy$ the *partial differentials*. In accordance with the arithmetical theory, this equation can only be regarded as a conveniently abridged form of the result obtained in the present discussion.

The expression $h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0}$ may be spoken of as the *first differential* of $f(x, y)$ at (x_0, y_0), and may be denoted by $\delta^{(1)}f$.

The theorem obtained may be stated as follows:

That the increment of a function $f(x_0, y_0)$ when x_0, y_0 are changed into x_0+h, y_0+k may be $h\frac{\partial f(x_0, y_0)}{\partial x_0} + k\frac{\partial f(x_0, y_0)}{\partial y_0} + h\rho + k\sigma$, where ρ, σ converge to zero, when h, k are indefinitely diminished, independently of the mode in which they are diminished, it is sufficient (1), that $\frac{\partial f(x_0, y_0)}{\partial x_0}$, $\frac{\partial f(x_0, y_0)}{\partial y_0}$ have definite finite values, and (2), that $\frac{f(x_0+h, y_0+k) - f(x_0, y_0+k)}{h}$ be a continuous function of (h, k) at the point $h = 0$, $k = 0$.

It will be observed that no assumption has been made that $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ have definite values except at the point (x_0, y_0) itself.

It follows from the definition that it is necessary for the existence of a total differential at (x_0, y_0) that the function be continuous with respect to (x, y), at (x_0, y_0).

If it be assumed that $\frac{\partial f}{\partial x}$ has a definite value at (x_0, y), for all values of y in some neighbourhood of y_0, the condition (2) may be expressed in the form that (a), $\frac{\partial f}{\partial x}$, for $x = x_0$, must be a continuous function of y, at $y = y_0$; and that (b), the point $h = 0$, $k = 0$ must be a point of uniform convergence of the function $\frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h}$, considered as a function of h, to its limit for $h \sim 0$, with k as a parameter, in accordance with the definition of such a point of uniform convergence given in § 305.

If it be assumed that $\frac{\partial f}{\partial x}$ exists, not merely at the point (x_0, y_0), but at all points in a sufficiently small two-dimensional neighbourhood of the point, the conditions contained in the theorem may be simplified. For we have, in that case,

$$\frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h} = \frac{\partial}{\partial x} f(x_0 + \theta h, y_0 + k),$$

where θ is such that $0 < \theta < 1$; and this expression converges to $\frac{\partial f(x_0, y_0)}{\partial x_0}$ provided $\frac{\partial f(x, y)}{\partial x}$ be continuous with respect to (x, y) at the point (x_0, y_0).

It has thus been proved that*, *in order that the increment of the function may be of the form given in the theorem above, it is sufficient* (1), *that* $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ *have definite values at the point* (x_0, y_0), *and* (2), *that one at least of these partial differential coefficients have definite values everywhere in a two-dimensional neighbourhood of* (x_0, y_0), *and be continuous at* (x_0, y_0), *with respect to the domain* (x, y).

310. Let it now be assumed that $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ exist, and are continuous functions, with respect to (x, y), at all points of an open domain D. Let D_1 be any perfect domain interior to D; at any point of D_1 we have

$$f(x + h, y + k) - f(x, y + k) = h \frac{\partial}{\partial x} f(x + \theta h, y + k),$$

---

* Thomae, *Einleitung in die Theorie der bestimmten Integrale*, p. 37.

where $0 < \theta < 1$, and h, k are such that $|h| < \zeta$, $|k| < \zeta$; the number ζ being so chosen that $\zeta\sqrt{2}$ is less than the distance between the boundaries of D_1 and D, so that the cell $(x-h, y-k; x+h, y+k)$ is certainly interior to D.

We have also

$$f(x, y+k) - f(x, y) = k\frac{\partial}{\partial y}f(x, y+\theta' k),$$

where $0 < \theta' < 1$.

Since $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ are continuous in the perfect domain D_1, they are uniformly continuous in that domain. Accordingly, a positive number ϵ $(< \zeta)$ can be so determined that, if η be a prescribed positive number,

$$\left|\frac{\partial}{\partial x}f(x+\theta h, y+k) - \frac{\partial}{\partial x}f(x, y)\right| < \eta,$$

and

$$\left|\frac{\partial}{\partial y}f(x, y+\theta' k) - \frac{\partial}{\partial y}f(x, y)\right| < \eta,$$

for all points (x, y), of D_1, provided $|h|$, $|k|$ are both less than ϵ.

The numbers η, ϵ converge to zero together. We have now the following theorem:

If the function $f(x, y)$ have partial differential coefficients that are continuous, with respect to the plane domain, at all points of an open domain D, then

$$f(x+h, y+k) = f(x, y) + h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} + hR + kR',$$

where (x, y), $(x+h, y+k)$ are any points of D, and R, R' tend to the limit zero, uniformly for all points (x, y) of any perfect domain contained in D, as h, k converge, in any manner, to zero.

EXAMPLES

1. Let* $f(x, y) = \sqrt{|xy|}$, where the positive value of the square root is to be taken. In this case $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ both exist at the point $(0, 0)$, and are both $= 0$. We have

$$\frac{f(h, k) - f(0, k)}{h} = \sqrt{\left|\frac{k}{h}\right|};$$

and this has different constant values for different constant values of k/h, and is therefore discontinuous at the point $h=0$, $k=0$. In this case the equation $f(h, k) = h\rho + k\sigma$, when ρ and σ converge to zero with h and k, cannot hold.

2. Let† $f(x, y) = x\sin(4\tan^{-1}y/x)$, for $x>0$; and $f(0, y)=0$, for all values of y. We find $\frac{\partial f(0, 0)}{\partial x}=0$, $\frac{\partial f(0, y)}{\partial x}=0$, and thus $\frac{\partial f(0, y)}{\partial x}$ is continuous with respect to y, at $(0, 0)$. Also, we find $\frac{\partial f(x, 0)}{\partial y}=4$, $\frac{\partial f(0, 0)}{\partial y}=0$, and therefore $\frac{\partial f(x, 0)}{\partial y}$ is discontinuous with regard

---

* Stolz, *Grundzüge*, vol. I, p. 133.

† Harnack's *Introduction to the Differential and Integral Calculus*, Cathcart's Translation, p. 93

to x, at $(0, 0)$. The value of $\frac{f(h, k)-f(0, k)}{h}$ is $\sin\left(4\tan^{-1}\frac{k}{h}\right)$, and this is discontinuous at $h=0$, $k=0$; hence the relation $df=\frac{\partial f}{\partial x}dx+\frac{\partial f}{\partial y}dy$ does not hold at the point $(0, 0)$.

HIGHER PARTIAL DIFFERENTIAL COEFFICIENTS

311. If the function $f(x, y)$ have the partial differential coefficient $\frac{\partial f}{\partial x}$, it may happen that, at the point (x_0, y_0), the function $\frac{\partial f}{\partial x}$ possesses a partial differential coefficient with respect to x. This is denoted by $\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2}$, and is spoken of as the *second partial differential coefficient* of $f(x, y)$ with respect to x, at the point x_0. The second partial differential coefficient $\frac{\partial^2 f(x_0, y_0)}{\partial y_0^2}$, with respect to y, is defined in a similar manner.

$\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2}$ is defined as the limit of $\frac{1}{h}\left\{\frac{\partial f(x_0+h, y_0)}{\partial x_0}-\frac{\partial f(x_0, y_0)}{\partial x_0}\right\}$; it being assumed that $\frac{\partial f(x_0, y_0)}{\partial x_0}$ has a definite value. It is not necessary for the existence of this limit, as a definite number, that $\frac{\partial f(x_0+h, y_0)}{\partial x_0}$, for $h \neq 0$, should have a definite value. When $\frac{\partial f(x_0+h, y_0)}{\partial x_0}$ has limits of indeterminancy

$$\overline{\frac{\partial f(x_0+h, y_0)}{\partial x_0}}, \quad \underline{\frac{\partial f(x_0+h, y_0)}{\partial x_0}},$$

it may happen that $\lim\limits_{h\sim 0}\frac{1}{h}\left\{\overline{\frac{\partial f(x_0+h, y_0)}{\partial x_0}}-\underline{\frac{\partial f(x_0+h, y_0)}{\partial x_0}}\right\}=0$, and that

$$\lim_{h\sim 0}\frac{1}{h}\left\{\overline{\underline{\frac{\partial f(x_0+h, y_0)}{\partial x_0}}}-\frac{\partial f(x_0, y_0)}{\partial x_0}\right\}$$

has a definite value.

Thus, in accordance with the definition of $\frac{\partial^2 f}{\partial x^2}$, this second partial differential coefficient may exist, as a definite number, at the point (x_0, y_0), but not at points in the neighbourhood; it is however assumed that $\frac{\partial f}{\partial x}$ exists at (x_0, y_0).

It may happen that, at the point (x_0, y_0), the function $\frac{\partial f}{\partial x}$ has a differential coefficient with respect to y: this may be denoted by $\frac{\partial}{\partial y_0}\left(\frac{\partial f}{\partial x_0}\right)$, or $\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$. Similarly, when $\frac{\partial f}{\partial y}$ has, at the point (x_0, y_0), a partial differential coefficient with respect to x, this is denoted by $\frac{\partial}{\partial x_0}\frac{\partial f}{\partial y_0}$, or $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0}$. These partial differential coefficients are said to be the mixed *partial differential coefficients* of the second order at (x_0, y_0) of $f(x, y)$, with respect to x and y, the order of differentiation being different in the two.

Under certain conditions which will have to be investigated, the two mixed partial differential coefficients of the second order, with respect to x and y, satisfy the relation

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} = \frac{\partial^2 f(x, y)}{\partial y \partial x},$$

which may be regarded as the fundamental theorem for partial differential coefficients of the second order.

The differential coefficient $\frac{\partial}{\partial x_0}\frac{\partial f}{\partial y_0}$, or $\frac{\partial^2 f}{\partial x_0 \partial y_0}$, is defined as the partial differential coefficient at (x_0, y_0), with respect to x_0, of $\frac{\partial f}{\partial y_0}$, it being assumed that $\frac{\partial f(x_0, y_0)}{\partial y_0}$ exists, as a definite number.

$$\text{Thus}\quad \frac{\partial^2 f}{\partial x_0 \partial y_0} = \lim_{h\sim 0}\frac{1}{h}\left[\lim_{k\sim 0}\frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0)}{k}\right.$$
$$\left. - \lim_{k\sim 0}\frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}\right].$$

For the existence of $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ as a definite number, it being assumed that $\frac{\partial f(x_0, y_0)}{\partial y_0}$ has a definite value, it is not necessary that $\frac{\partial f(x_0 + h, y_0)}{\partial y_0}$ should have a definite value for $h \neq 0$. $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ may exist as

$$\lim_{h\sim 0}\frac{1}{h}\left\{\overline{\frac{\partial f(x_0 + h, y_0)}{\partial y_0}} - \frac{\partial f(x_0, y_0)}{\partial y_0}\right\}.$$

When the differential coefficient $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ exists, it is equal to the repeated limit

$$\lim_{h\sim 0}\lim_{k\sim 0}\frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0) - f(x_0, y_0 + k) + f(x_0, y_0)}{hk}.$$

Conversely, if this repeated limit has a definite value, and if also

$$\lim_{k\sim 0}\frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}$$

exists, then $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ exists, and is equal to the repeated limit.

It is however possible that the repeated limit may exist, and yet that $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ may not exist, if the condition that $\frac{\partial f}{\partial y_0}$ exists be not satisfied.

For example* let $f(x, y) = \chi_1(x) + \chi_2(y)$; where $\chi_1(x)$, $\chi_2(y)$ are non-differentiable functions of x and y respectively. In this case

$$\frac{\partial f(x_0 + h, y_0)}{\partial y_0}, \quad \frac{\partial f(x_0, y_0)}{\partial y_0}$$

do not exist, and thus $\frac{\partial^2 f}{\partial x_0 \partial y_0}$ does not exist, but

$$\lim_{h \sim 0} \lim_{k \sim 0} \frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0) - f(x_0, y_0 + k) + f(x_0, y_0)}{hk}$$

exists, and is equal to zero.

It would be possible to define $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0}$ as the value of this last repeated limit, when it exists, in which case $\frac{\partial f(x_0, y_0)}{\partial y_0}$ would not necessarily exist. It is however more convenient to restrict the definition, as has been done above, to apply only to the case in which $\frac{\partial f(x_0, y_0)}{\partial y_0}$ exists, as a definite number: accordingly this definition will be adopted.

It will also be assumed that $\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2}$ exists, only when $\frac{\partial f(x_0, y_0)}{\partial x_0}$ exists. A similarly restricted definition of $\frac{\partial^2 f}{\partial y_0 \partial x_0}$ will be employed.

312. Denoting

$$f(x_0 + h, y_0 + k) - f(x_0 + h, y_0) - f(x_0, y_0 + k) + f(x_0, y_0)$$

by $F(h, k)$, the condition that $\frac{\partial^2 f}{\partial x_0 \partial y_0} = \frac{\partial^2 f}{\partial y_0 \partial x_0}$ holds is identical with the condition that the two repeated limits of $\frac{F(h, k)}{hk}$, for $h \sim 0$, $k \sim 0$, should both exist, and have the same finite, or infinite, value; it being assumed that $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial x}$ both exist at the point (x_0, y_0). The necessary and sufficient conditions for the equality of the two partial differential coefficients may be obtained by applying the conditions given in either of the theorems in § 304, and § 306, to the function $F(h, k)/hk$. It is however convenient, for application in particular cases, to possess sufficient conditions, of a simple character, relating to the partial differential coefficients in the neighbourhood of the point (x_0, y_0).

---

* See Hobson, *Proc. Lond. Math. Soc.* (2), vol. v (1907), p. 233.

The following theorem will be established:

If (1), $\frac{\partial^2 f(x, y)}{\partial y \partial x}$ *exists, and is finite, at all points in a plane neighbourhood* (ϵ) *of the point* (x_0, y_0), *except that its existence at the point* (x_0, y_0) *itself is not assumed, and if* (2), $\frac{\partial^2 f(x, y)}{\partial y \partial x}$ *has a unique double limit* A, *at the point* (x_0, y_0), *which is either finite, or infinite with a fixed sign, and if* (3), $\frac{\partial f(x_0, y_0)}{\partial x_0}$, $\frac{\partial f(x_0, y_0)}{\partial y_0}$ *both exist, and are finite, then* $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0}$, $\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$ *both exist, and have the same value* A.

In particular, the conditions of the theorem are satisfied if $\frac{\partial^2 f}{\partial y \partial x}$ *exists everywhere, without exception, in a neighbourhood of* (x_0, y_0), *and is continuous at* (x_0, y_0), *and if* $\frac{\partial f}{\partial y}$ *exists at* (x_0, y_0).

In the first place, we observe that, as a consequence of (1), $\frac{\partial f(x, y)}{\partial x}$ exists everywhere in the neighbourhood (ϵ) of (x_0, y_0), its existence at the point (x_0, y_0) being assumed, in accordance with (3). Moreover $\frac{\partial f(x, y)}{\partial x}$ must be continuous with respect to y, except possibly at the point (x_0, y_0) itself. It also follows that $f(x, y)$ is continuous with respect to x, at every point of the neighbourhood of (x_0, y_0), including the point (x_0, y_0) itself, on account of (3).

Let us suppose that h, k are both positive; the other three quadrants may then be considered separately, in the same manner.

Since $\frac{\partial f(x_0 + h, y)}{\partial x}$, where $0 < h < \epsilon$, is continuous with respect to y, in the closed interval $y_0 \leq y \leq y_0 + k$, where $k < \epsilon$, and since it possesses a differential coefficient with respect to y, at every interior point of that interval of y, we have, by the mean value theorem (§ 262),

$$\frac{\partial f(x_0 + h, y_0 + k)}{\partial x} - \frac{\partial f(x_0 + h, y_0)}{\partial x} = k \frac{\partial^2 f(x_0 + h, y_0 + \theta k)}{\partial y \partial x},$$

where θ is some number such that $0 < \theta < 1$. The expression on the right-hand side is, by (2), equal to $k[A - \alpha(h, k)]$, if A be finite, where $|\alpha(h, k)| < \epsilon_1$, an arbitrarily chosen positive number, provided h, k are each less than some fixed positive number η, dependent on ϵ_1. In case $A = +\infty$, the corresponding result is that

$$\left| \frac{\partial f(x_0 + h, y_0 + k)}{\partial x} - \frac{\partial f(x_0 + h, y_0)}{\partial x} \right| > kN,$$

for $0 < h < \eta$, $0 < k < \eta$, where N is an arbitrary positive number, and η depends on N.

Since the function $f(x, y_0 + k) - f(x, y_0)$ is a continuous function of x, in the closed interval $(x_0, x_0 + h)$, and since it possesses a differential coefficient with respect to x, at all points x in that interval, we have, by applying the mean value theorem,

$$F(h, k) = h\left\{\frac{\partial f(x_0 + \theta' h, y_0 + k)}{\partial x} - \frac{\partial f(x_0 + \theta' h, y_0)}{\partial x}\right\},$$

where θ' is a number such that $0 < \theta' < 1$.

From the two results obtained, we have $F(h, k)/hk = A + \alpha'(h, k)$, when A is finite, where $|\alpha'(h, k)| < \epsilon_1$, provided h, k are both less than η.

In case $A = -\infty$, $|F(h, k)/hk| > N$, if h, k are both less than η.

The case $A = -\infty$ can be treated similarly. In either case $F(h, k)/hk$ has the double limit A. Since $\frac{\partial f}{\partial x_0}, \frac{\partial f}{\partial y_0}$ both exist at (x_0, y_0), as finite numbers, it follows that $\frac{\partial^2 f}{\partial x_0 \partial y_0}, \frac{\partial^2 f}{\partial y_0 \partial x_0}$ both exist, and have the value A.

The sufficient conditions in the foregoing theorem are somewhat simpler than those stated by Schwarz*, who assumed the additional condition that $\frac{\partial f(x, y_0)}{\partial y_0}$ exists, and is finite, for all values of x in the neighbourhood of $x = x_0$, for the constant value y_0, of y.

A somewhat different set of sufficient conditions has been given† by W. H. Young, in the following theorem:

If (1), $\frac{\partial^2 f}{\partial y \partial x}$ *exists, and is finite, at all points of the neighbourhood* (ϵ) *of the point* (x_0, y_0), *except those points for which* $x = x_0$ *or* $y = y_0$, *its existence on these straight lines not being assumed, and if* (2), *these values of* $\frac{\partial^2 f}{\partial y \partial x}$ *have a unique double limit* A (*finite, or infinite with fixed sign*), *and if* (3), $\frac{\partial f}{\partial x}$ *exists on the two straight lines* $x = x_0$, $y = y_0$ *in the neighbourhood* (ϵ) *of* (x_0, y_0), *including that point itself, and is continuous with respect to y, when* $y = y_0$, $x \neq x_0$, *and if* (4), $\frac{\partial f}{\partial y}$ *exists when* $y = y_0$, *for all values of x in the neighbourhood* (ϵ) *of* x_0, *including* x_0 *itself, then* $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0}$, $\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$ *both exist, and have the value* A.

---

* *Gesammelte Abh.* vol. II (1890), p. 275, see also Peano, *Mathesis*, vol. X, p. 153, and also Stolz, *Grundzüge d. Diff. Rech.* vol. I, p. 147. Sufficient conditions were also investigated by Hobson, *Proc. Lond. Math Soc.* (2), vol. V (1907), p. 225. See also Fubini, *Gior. di Battaglini*, vol. XXXVIII (1900), p. 72, and Dini, *Lezioni di Analysi infinitesimale*, vol. I (1907), p. 164.

† *Proc. R.S.E.* vol. XXIX (1908–9), p. 136.

Only a slight modification of the proof of the last theorem is required in order to prove this theorem.

From (1) it follows that $\frac{\partial f}{\partial x}$ exists, and is a continuous function of y, when $x \neq x_0$, $y \neq y_0$; this, combined with (3), shews that $\frac{\partial f}{\partial x}$ exists everywhere in the neighbourhood (ϵ) of (x_0, y_0), and that it is everywhere continuous with respect to y, except possibly at the point (x_0, y_0). It follows that $f(x, y)$ is everywhere continuous with respect to x. The proof then proceeds as before.

313. *If* $\frac{\partial f(x, y)}{\partial x}$, $\frac{\partial f(x, y)}{\partial y}$ *have total differentials*[*] *at the point* (x_0, y_0), *then*

$$\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0} = \frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}.$$

In accordance with the condition stated, the two functions $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ are continuous with respect to (x, y) at the point (x_0, y_0), and they are accordingly exist in a neighbourhood of (x_0, y_0). Hence $f(x, y)$ is continuous both with respect to x, and with respect to y. We have

$$\frac{\partial f(x_0 + h, y_0 + k)}{\partial x} - \frac{\partial f(x_0, y_0)}{\partial x_0} = h\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2} + k\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0} + h\rho + k\sigma,$$

where $|\rho|$, $|\sigma|$ converge to zero as h, k do so in any manner.

We have

$$F(h, k) = \{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0)\} - \{f(x_0, y_0 + k) - f(x_0, y_0)\}$$
$$= h\left\{\frac{\partial f(x_0 + \theta h, y_0 + k)}{\partial x} - \frac{\partial f(x_0 + \theta h, y_0)}{\partial x}\right\},$$

where θ is such that $0 < \theta < 1$. From this, we see that

$$F(h, k) = h\left\{\theta h\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2} + k\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0} + \theta h\rho' + k\sigma' - \theta h\frac{\partial^2 f(x_0, y_0)}{\partial x_0^2} - \theta h\rho''\right\},$$

where $|\rho'|$, $|\sigma'|$, $|\rho''|$ converge to zero, as $h \sim 0$, $k \sim 0$.

We now have

$$\frac{F(h, k)}{hk} = \frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0} + \theta\frac{h}{k}\rho' + \sigma' - \theta\frac{h}{k}\rho'';$$

hence, if h and k converge to zero in any manner such that k/h is greater than a fixed positive number ϵ, $F(h, k)/hk$ converges to $\frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$.

---

[*] See W. H. Young, *loc. cit.*

Similarly, it can be shewn that $F(h, k)/hk$ converges to $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0}$, if h and k converge to zero so that $h/k > \epsilon$.

By letting h and k converge to zero, so that both h/k and k/h are $> \epsilon$, we see that $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0} = \frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$.

It has been shewn, in § 309, that the conditions of the above theorem are satisfied if the four partial differential coefficients of the second order all exist at the point (x_0, y_0), and if $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$ exist in a neighbourhood of that point and are continuous with respect to (x, y), at the point (x_0, y_0). The following theorem has thus been proved:

If the four partial differential coefficients of $f(x, y)$ exist at the point (x_0, y_0), and if $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$ exist in a neighbourhood of (x_0, y_0), and are continuous relative to (x, y), at the point (x_0, y_0), then $\frac{\partial^2 f(x_0, y_0)}{\partial x_0 \partial y_0} = \frac{\partial^2 f(x_0, y_0)}{\partial y_0 \partial x_0}$.

314. The partial differential coefficients of higher order n of a function $f(x, y)$ are of the form $\frac{\partial^n f(x, y)}{\partial x^p \partial y^q \partial x^r \ldots \partial x^k \partial y^l}$, where $p, q, r, \ldots l$ are positive integers, including zero, such that $p + q + r + \ldots + l = n$. Here, f is first differentiated l times with respect to y, then k times with respect to x, and so on. The total number of possible partial differential coefficients of order n is 2^n; the number of those in which r differentiations with respect to x, and $n - r$ with respect to y are involved is $\frac{n!}{r!\,(n-r)!}$.

Sufficient conditions for the existence of all the partial differential coefficients of order n may be obtained by extending the theorem of § 311, which refers to the case $n = 2$. The following criteria*, which can be proved by induction, will be sufficient for the purpose:

If the $n - 1$ differential coefficients $\frac{\partial^n f}{\partial x^{n-1} \partial y}$, $\frac{\partial^n f}{\partial x^{n-2} \partial y^2}$, $\ldots$ $\frac{\partial^n f}{\partial x \partial y^{n-1}}$ have definite finite values for all points in a two-dimensional neighbourhood of the point (x_0, y_0), and are continuous at the point (x_0, y_0), with respect to (x, y), then all the other mixed partial differential coefficients of order n exist at the point (x_0, y_0); and each one of them has the same value at the point as that one of those given above in which the same number of differentiations with respect to x, and with respect to y, occurs, as in the one considered.

---

* See Stolz, *Grundzüge*, vol. I, p. 153.

EXAMPLE

Let* the function $f(x, y)$ be defined by $f(x, y) = xy\frac{x^2-y^2}{x^2+y^2}$ for all values of x and y, except when $x=0$, $y=0$; for which $f(0, 0)=0$. At the point $(0, 0)$ the partial differential coefficients $\frac{\partial^2 f}{\partial x\partial y}$, $\frac{\partial^2 f}{\partial y\partial x}$ both exist, and have different finite values.

The function $f(x, y)$ is continuous at the point $(0, 0)$; for, putting $x = r\cos\theta$, $y = r\sin\theta$, the function becomes $\frac{1}{4}r^2\sin 4\theta$; and this is numerically less than ϵ, provided $r < 2\sqrt{\epsilon}$.

We find $\frac{\partial f(x, y)}{\partial x} = y\left\{\frac{x^2-y^2}{x^2+y^2} + \frac{4x^2y^2}{(x^2+y^2)^2}\right\}$, at any point except $(0, 0)$; at the point $(0, 0)$, $\frac{\partial f}{\partial x}$ is $\lim_{x\sim 0}\frac{f(x, 0)-f(0, 0)}{x}$, which is $= 0$.

The value of $\frac{\partial f(0, y)}{\partial x}$ is $-y$, and that of $\frac{\partial f(x, 0)}{\partial y}$ is x.

We then find
$$\frac{\partial^2 f(0, 0)}{\partial y\partial x} = \lim_{y=0}\frac{1}{y}\left\{\frac{\partial f(0, y)}{\partial x} - \frac{\partial f(0, 0)}{\partial x}\right\} = -1,$$
and
$$\frac{\partial^2 f(0, 0)}{\partial x\partial y} = \lim_{x=0}\frac{1}{x}\left\{\frac{\partial f(x, 0)}{\partial y} - \frac{\partial f(0, 0)}{\partial y}\right\} = 1.$$

The value of $\frac{\partial^2 f(x, y)}{\partial x\partial y}$, as also that of $\frac{\partial^2 f(x, y)}{\partial y\partial x}$, is $\frac{x^2-y^2}{x^2+y^2}$ [illegible] at every point except $(0, 0)$. This value is $\cos 2\theta(1+2\sin^2 2\theta)$, which is constant for constant θ, [illegible] but has different values for different values of θ; and thus the partial differential coefficients are discontinuous at the point $(0, 0)$. The conditions of the theorem of [illegible] and the sufficient conditions for the equality of $\frac{\partial^2 f}{\partial x\partial y}$ and $\frac{\partial^2 f}{\partial y\partial x}$ are therefore not satisfied for the point $(0, 0)$.

315. Let $\phi(h)$ be a function of h, such that $\phi(0) = 0$, and such that in an open neighbourhood of the point $h = 0$, the first $n-1$ differential coefficients of $\phi(h)$ all exist, and all have the value zero at $h = 0$.

We have then, from the theorem of § 263,
$$\frac{\phi(h)}{h^n} = \frac{\phi'(h_1)}{nh_1^{n-1}} = \frac{\phi''(h_2)}{n(n-1)h_2^{n-2}} = \dots = \frac{\phi^{(n-1)}(h_{n-1})}{n!\,h_{n-1}},$$
where $0 < |h_{n-1}| < |h_{n-2}| < \dots < |h_1| < |h|$.

Now let $\phi(h) = F(h, x, y)$; and let us assume that, for all points (x, y) in a given open plane domain D, and for all values of h in a given open neighbourhood of $h = 0$, the first $n-1$ partial differential coefficients of $F(h, x, y)$ with respect to h exist, and are all zero, when $h = 0$; and also that $F(0, x, y) = 0$. We have then
$$\frac{F(h, x, y)}{h^n} = \frac{1}{n!\,h_{n-1}}\frac{\partial^{n-1}F(h_{n-1}, x, y)}{\partial h_{n-1}^{n-1}},$$
where h_{n-1} is such that $0 < |h_{n-1}| < |h|$; the value of h_{n-1} depending on the values of h, x, and y.

---

* Peano, *Calc. Diff.* p. 174.

If $\frac{\partial^{n-1} F(h, x, y)}{\partial h^{n-1}}$ have partial differential coefficients with respect to h, x, y, that are continuous in the open domain for which (x, y) is in D, and h in the given open interval, then, employing the theorem of § 310, as extended to the case of a function of the three variables h, x, y, we see that

$$\frac{\partial^{n-1} F(h_{n-1}, x, y)}{\partial h_{n-1}^{n-1}} = h_{n-1}\left[\left\{\frac{\partial^n F(h, x, y)}{\partial h^n}\right\}_{h=0} + R\right];$$

where R converges to zero, as h_{n-1} converges to zero, uniformly for all points (x, y) in a perfect domain D_1, contained in D, and for all points h_{n-1} that correspond to a point h in a closed neighbourhood of $h = 0$, interior to the given open neighbourhood.

Therefore, if $\bar{h}$ be sufficiently small, and $|h| \leq \bar{h}$, we have

$$\left|\frac{F(h, x, y)}{h^n} - \frac{1}{n!}\left\{\frac{\partial^n F(h, x, y)}{\partial h^n}\right\}_{h=0}\right| < \epsilon,$$

for all points (x, y) in D_1.

Now let

$$F(h, x, y) = f(x + h, y) - f(x, y) - h\frac{\partial f(x, y)}{\partial x} - \frac{h^2}{2!}\frac{\partial^2 f(x, y)}{\partial x^2} - \dots - \frac{h^{n-1}}{(n-1)!}\frac{\partial^{n-1} f(x, y)}{\partial x^{n-1}};$$

this satisfies the conditions that F and its first $n - 1$ differential coefficients with respect to h all vanish when $h = 0$; we have therefore, provided $f(x, y)$ and its partial differential coefficients of the first n orders are continuous in D,

$$f(x + h, y) = f(x, y) + h\frac{\partial f(x, y)}{\partial x} + \frac{h^2}{2!}\frac{\partial^2 f(x, y)}{\partial x^2} + \dots + \frac{h^n}{n!}\left\{\frac{\partial^n f(x, y)}{\partial x^n} + R'\right\},$$

where R' converges to zero, as $h \sim 0$, uniformly for all points (x, y) in the perfect domain D_1, interior to D.

Similarly, we find that

$$f(x, y + k) = f(x, y) + k\frac{\partial f(x, y)}{\partial y} + \dots + \frac{k^n}{n!}\left\{\frac{\partial^n f}{\partial y^n} + R''\right\},$$

where R'' converges to zero, as $k \sim 0$, uniformly in D_1.

In the first of these equations, we can write $y + k$ for y, where k is so small that $(x, y + k)$ is interior to D, for all points (x, y), of D_1. We have then

$$f(x + h, y + k) - f(x, y + k)$$
$$= h\frac{\partial f(x, y + k)}{\partial x} + \dots + \frac{h^r}{r!}\frac{\partial^r f(x, y + k)}{\partial x^r} + \dots + \frac{h^n}{n!}\left\{\frac{\partial^n f(x, y + k)}{\partial x^n} + R'''\right\},$$

where R''' converges to 0, as $h \sim 0$, uniformly in D_1, and uniformly with respect to k, for all values which do not numerically exceed a fixed number.

Since all the partial differential coefficients of the first n orders are continuous in D, we have, provided $|h|$, $|k|$ are so small that, when (x, y) is in D_1, the points $(x+h, y+k)$ are all interior to D,

$$\frac{\partial^r f(x, y+k)}{\partial x^r} = \frac{\partial^r f(x, y)}{\partial x^r} + k\frac{\partial^{r+1} f(x, y)}{\partial y \partial x^r} + \dots + \frac{k^{n-r}}{(n-r)!}\frac{\partial^n f}{\partial y^{n-r}\partial x^r} + R_r,$$

where R_r converges uniformly to zero, as $k \sim 0$.

Substituting this expression, for $r = 1, 2, 3, \dots n$, in the expression for $f(x+h, y+k) - f(x, y+k)$, and adding the expression for

$$f(x, y+k) - f(x, y),$$

we obtain the formula

$$f(x+h, y+k) - f(x, y) = h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} + \frac{1}{2!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 f$$

$$+ \dots + \frac{1}{r!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^r f$$

$$+ \dots + \frac{1}{n!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f + \bar{R},$$

where $\bar{R} = a_{n1}h^n + a_{n2}h^{n-1}k + \dots + a_{nn}k^n$, and $a_{n1}, a_{n2}, \dots a_{nn}$ all converge to zero, as h, k converge in any manner to zero, uniformly for all points (x, y) of the perfect domain D_1, interior to D.

$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^r f$ is used to denote

$$h^r\frac{\partial^r f}{\partial x^r} + rh^{r-1}k\frac{\partial^r f}{\partial y \partial x^{r-1}} + \frac{r(r-1)}{2!}h^{r-2}k^2\frac{\partial^r f}{\partial y^2\partial x^{r-2}} + \dots + k^r\frac{\partial^r f}{\partial y^r}.$$

That $\frac{\partial^n f}{\partial x^r \partial y^{n-r}} = \frac{\partial^n f}{\partial y^{n-r}\partial x^r}$ follows from the condition of the continuity of the partial differential coefficients of the first n orders in the domain D. That this is the case is verified by the fact that the expansion obtained above may also be found by exchanging the parts which x and y play in the process by which the expression was obtained.

It has thus been shewn that:

If $f(x, y)$, and all its first n partial differential coefficients with respect to x and y are continuous in an open domain D, of (x, y), then for any pair of points $(x+h, y+k)$, (x, y), in D,

$$f(x+h, y+k) = f(x, y) + \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)f + \frac{1}{2!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 f$$

$$+ \dots + \frac{1}{n!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f + \bar{R},$$

where $\bar{R} = a_{n1}h^n + a_{n2}h^{n-1}k + \dots + a_{nn}k^n$; and $a_{n1}, a_{n2}, \dots a_{nr}$ converge to zero, as h, k converge in any manner to zero, uniformly for all points (x, y) in a perfect domain interior to D.

This theorem is the generalization of that of § 310, which corresponds to the case $n = 1$.

The expression $\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^r f(x, y)$ is called the rth differential of $f(x, y)$, and may be denoted by $\delta^{(r)} f(x, y)$; the theorem may then be written in the form

$$f(x+h, y+k) - f(x, y)$$
$$= \delta^{(1)} f(x, y) + \frac{1}{2!}\delta^{(2)} f(x, y) + \dots + \frac{1}{n!}\{\delta^{(n)} f(x, y) + \bar{R}\}.$$

If (x, y) be a fixed point (x_0, y_0), and

$$f(x_0+h, y_0+k) - f(x_0, y_0)$$
$$= \delta^{(1)} f(x_0, y_0) + \frac{1}{2!}\delta^{(2)} f(x_0, y_0) + \dots + \frac{1}{n!}\{\delta^{(n)} f(x_0, y_0) + \bar{R}\},$$

where $\bar{R}$ has the form and properties given above, the function $f(x, y)$ is said to possess an nth total differential at the point (x_0, y_0). In accordance with the above theorem this nth total differential certainly exists in case there exists a plane neighbourhood of (x_0, y_0) in which all the partial differential coefficients of the first n orders exist, and are continuous. But, as has been shewn in § 309, for the case $n = 1$, less stringent conditions than these are sufficient to ensure the existence of the nth total differential at the point (x_0, y_0). It has been shewn* by W. H. Young that, if $f(x, y)$ have an $(n - 1)$th total differential at (x_0, y_0), while the partial differential coefficients of order $n - 1$ all exist, and are independent of the order of differentiation, in a closed neighbourhood of the point (x_0, y_0), and further, if they all have first differentials at that point, then the function has an nth total differential at (x_0, y_0), so that the above expansion theorem holds for the number n. The subject has also been treated† by Rademacher.

FUNCTIONS DEFINED IMPLICITLY

316. A function y, of the variable x, is said to be defined *implicitly* when there is given a functional relation between the variables of the form $F(x, y) = 0$; provided that this relation suffices to determine uniquely, in some domain of x, the corresponding values of y.

The first general theorem as to the existence of the function $y = \phi(x)$, determined implicitly by a relation $F(x, y) = 0$, was given by Cauchy, for the case in which $F(x, y)$ can, in a neighbourhood of a fixed point (α, β) at which $F(\alpha, \beta) = 0$, be represented by a convergent series proceeding by powers of $x - \alpha$, $y - \beta$.

The theorem was freed by Dini‡ from the restriction that $F(x, y)$ must be analytic, and his theorem may be stated as follows:

---

* *Proc. Lond. Math. Soc.* (2), vol. VII (1908), p. 171.

† *Math. Annalen*, vol. LXXIX (1919), p. 340.

‡ *Analysi infinitesimale*, vol. I, p. 162.

If $F(x, y)$ is continuous in a neighbourhood of [illegible] *$\frac{\partial F}{\partial y}$ exists, and is continuous in this neighbourhood, and does not vanish at the point (α, β); then, corresponding to an arbitrarily small positive* [illegible] *another such number ϵ can be so determined that, to each value of* [illegible] *interval $(\alpha - \epsilon, \alpha + \epsilon)$, there corresponds a* [illegible] *interval $(\beta - \delta, \beta + \delta)$, so that these corresponding values* [illegible] *satisfy the condition $F(x, y) = 0$. The function* [illegible] *tinuous in the open interval $(\alpha - \epsilon, \alpha + \epsilon)$. Moreover,* [illegible] *is satisfied that $\frac{\partial F}{\partial x}$ exists, and is continuous in* [illegible] *the function $\phi(x)$ has a continuous differential* [illegible] *$(\alpha - \epsilon, \alpha + \epsilon)$.*

Since $\frac{\partial F}{\partial y}$ is continuous in a neighbourhood of [illegible] vanish at that point, a positive number h can be so determined that [illegible]

$$\alpha - h < x < \alpha + h,\ \beta - h < y < \beta + h,$$

$\frac{\partial F}{\partial y}$ does not vanish at (x, y), and has the same sign as at [illegible] assumed that the sign is positive. The number h can be so chosen that the closed neighbourhood (h) of the point (α, β) is interior to the neighbourhood in which F and $\frac{\partial F}{\partial y}$ are continuous.

Since $\frac{\partial F}{\partial y}$ is positive for $\beta - h < y < \beta + h$, [illegible] when $x = \alpha$, $y = \beta$, $F(\alpha, y)$ must be negative for [illegible] positive for $\beta < y \leq \beta + h$. Since $F(\alpha, \beta - h)$ is negative, a positive number k_1 ($\leq h$) can be so determined that $F(x, \beta - h) < 0$, for

$$\alpha - k_1 < x < \alpha + k_1.$$

Similarly, a positive number k_2 ($\leq h$) can be so determined that

$$F(x, \beta + h) > 0,$$

for $\alpha - k_2 < x < \alpha + k_2$. If k be the smaller of the two numbers k_1, k_2, we see that, for $\alpha - k < x < \alpha + k$, $F(x, \beta + h) > 0$, and $F(x, \beta - h) < 0$.

Since $\frac{\partial F}{\partial y} > 0$, for $\alpha - k < x < \alpha + k$, there must be one value of y, and only one such that $|y - \beta| < h$, at which $F(x, y)$ vanishes, for each value of x such that $|x - \alpha| < k$. Thus the function $y = \phi(x)$ is determined for $|x - \alpha| < k$. When h and k have been determined, we may take $\delta \leq h$; then a value of ϵ, $\leq k$, can be determined, corresponding to δ.

If $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$ are both continuous, when $|x - \alpha| \leq h'$, $|y - \beta| \leq k'$, where $h' < h$, $k' < k$, the difference ΔF of the values of the function F at any two points $(x - \Delta x, y - \Delta y)$, (x, y) in that neighbourhood of (α, β), is given by

$$\Delta F = \left(\frac{\partial F}{\partial x} + \rho\right) \Delta x + \left(\frac{\partial F}{\partial y} + \sigma\right) \Delta y,$$

where ρ, σ are continuous functions that converge to zero, as Δx, Δy do so, uniformly for all the points (x, y).

When $y = \phi(x)$, $y + \Delta y = \phi(x + \Delta x)$, we have $\Delta F = 0$; thus

$$\frac{\Delta y}{\Delta x} = -\left(\frac{\partial F}{\partial x} + \rho\right) \Big/ \left(\frac{\partial F}{\partial y} + \sigma\right),$$

and the expression on the right-hand side converges uniformly to

$$-\frac{\partial F}{\partial x} \Big/ \frac{\partial F}{\partial y}.$$

Thus the differential coefficient $\phi'(x)$ exists, and is continuous in the open interval $\alpha - k < x < \alpha + k$.

In case the function $F(x, y)$ has the form $f(y) - x$, we have the case in which the function $\phi(x)$ is determined as the inverse of the function $x = f(y)$. The theorem then takes the following form:

If $f(y)$ and its differential coefficient $f'(y)$ are continuous in a neighbourhood of $y = \beta$, and if $f'(y)$ does not vanish when $y = \beta$, then, corresponding to a sufficiently small positive number δ, another such number ϵ can be so determined that, to each value of x in the open interval $(\alpha - \epsilon, \alpha + \epsilon)$, there corresponds a unique value of y in the open interval $(\beta - \delta, \beta + \delta)$, such that the corresponding values of x and y satisfy the relation $x = f(y)$. The function $y = \phi(x)$, so determined, is continuous, and has a continuous differential coefficient in the open interval $(\alpha - \epsilon, \alpha + \epsilon)$.

316¹. The following more general theorem was given* by W. H. Young:

Let $F(x, y)$ be continuous in a neighbourhood of the point (a, b), at which $F(a, b) = 0$, both with respect to x and with respect to y, and let it be assumed that, in that neighbourhood, $\frac{\partial F}{\partial y}$ exists and is everywhere finite, and that it does not vanish at the point (a, b); then (1), there exists a function $y = \phi(x)$ such that $b = \phi(a)$, and such that, when this value of y is substituted in $F(x, y)$, the condition $F(x, y) = 0$ is satisfied everywhere in a certain neighbourhood of (a, b). Also (2), if it be the case that, for each fixed value of x, $F(x, y)$ is, in some closed neighbourhood of (a, b), a monotone function of y which is not constant in any interval, then the function $y = \phi(x)$ is unique. Further (3), if $F(x, y)$ be a continuous function of (x, y) in the closed neighbourhood of (a, b), then the unique function $\phi(x)$ is a continuous function

---

* *Proc. Lond. Math. Soc.* (2), vol. VII (1909), p. 404.

of x. Lastly (4), if $F(x, y)$ possesses at (a, b) a first differential, then the function $\phi(x)$ has, at $x = a$, a first differential coefficient $\phi'(a)$, given by

$$\frac{\partial F(a, b)}{\partial a} + \phi'(a)\frac{\partial F(a, b)}{\partial b} = 0.$$

We may assume that $\frac{\partial F(a, b)}{\partial b}$ is positive; then a positive number k can be so determined that $F(a, y)$ is positive when $b < y \leq b + k$, and that it is negative when $b - k \leq y < b$. We may choose k so small that the points $(a, b - k)$, $(a, b + k)$ are both interior to the neighbourhood of (x, y) in which $F(x, y)$ is continuous with respect to x and with respect to y, and for which $\frac{\partial F}{\partial y}$ exists and is finite. A rectangle of which the corners are $(a - h, b - k)$, $(a - h, b + k)$, $(a + h, b - k)$, $(a + h, b + k)$ may then be constructed, so as to be interior to this neighbourhood, and to be such that $F(x, b - k) > 0$, for $a - h \leq x \leq a + h$, and that $F(x, b + k) < 0$, for $a - h \leq x \leq a + h$, since $F(x, y)$ is continuous with respect to x. The rectangle so constructed is such that the conditions (1) are satisfied. For on each ordinate parallel to the y-axis, and contained in the rectangle $F(x, y)$ has both positive and negative values, and from the continuity of the function with respect to y, it follows that the function vanishes at one or more points which form a closed set. Taking on each ordinate the upper or lower boundary of this closed set we obtain a function $\phi(x)$ which satisfies the condition (1).

In condition (2), $F(x, y)$ is monotone along each ordinate and is not constant in any interval, the function accordingly vanishes at only one point on that ordinate, and the function $\phi(x)$ is then unique.

If further (3), $F(x, y)$ is continuous with respect to (x, y) in the rectangle, the plane set of its zeros is a closed set. Taking a sequence $\{x_n\}$ of values of x, which converges to the value x, the set of zeros (x_n, y_n) of $F(x, y)$ converges to the zero (x, y), so that, for every such value of x, $\phi(x)$ is the limit of $\phi(x_n)$, and thus $\phi(x)$ is a continuous function.

If lastly (4), $F(x, y)$ has a first differential at (a, b), there is a closed neighbourhood of the point such that

$$F(x, y) = (x - a)\left[\frac{\partial F(a, b)}{\partial a} + \rho\right] + (y - b)\left[\frac{\partial F(a, b)}{\partial b} + \sigma\right],$$

where ρ and σ converge to zero with $x - a$ and $y - b$. Taking (x, y) to be a zero of F, we have

$$\frac{y - b}{x - a} = -\frac{\frac{\partial F(a, b)}{\partial a} + \rho}{\frac{\partial F(a, b)}{\partial b} + \sigma}.$$

As x approaches a in any manner, y being continuous has the limit b, and thus ρ and σ have the limits zero, we have then

$$\frac{\partial F(a,b)}{\partial a} + \phi'(a)\frac{\partial F(a,b)}{\partial b} = 0.$$

In the theorem either of the following conditions may be substituted for (2):

(2)′, *if* $\frac{\partial F}{\partial y}$ *exists throughout a closed neighbourhood of* (a, b), *and has at no point therein the value zero, then the function* $y = \phi(x)$ *is unique.*

(2)″, *if* $\frac{\partial F}{\partial y}$ *exists throughout a closed neighbourhood of* (a, b), *and is continuous at the point* (a, b) *with respect to* (x, y), *then the function* $y = \phi(x)$ *is unique.*

For, on an ordinate in the rectangle, $\frac{\partial F}{\partial y}$ must everywhere have the same sign, since, if it had both positive and negative values, it must take every value between its upper and lower boundaries (see § 278), and thus it must at least at one point have the value zero, which is not the case. It then follows that, for each value of x, $F(x, y)$ is a monotone function of y, and thus the condition in (2) is satisfied. Therefore (2)′ may be substituted for (2).

Again, if the condition in (2)″ is satisfied, it follows that, in some closed neighbourhood of (a, b), $\frac{\partial F}{\partial y}$ is everywhere different from zero; and thus the condition in (2)′ is satisfied.

317. With a view to the extension of Dini's theorem of § 316, the following lemmas* will be required:

If $f(x_1, x_2, \ldots x_p)$ *is continuous in an open p-dimensional set O, and is either* (1), *positive, or* (2), *different from zero, at every point of a closed set G contained in O, then a neighbourhood of G, open or closed, can be so determined that the function has the same property* (1), *or* (2), *in this neighbourhood, as in G itself.*

Let K be a closed set that contains G, and is contained in O. Those points of K (if any) for which, in case (1), $f \leqq 0$, or in case (2), $f = 0$, form a closed set κ; this follows from the continuity of f. The two closed sets G, κ have no point in common, and therefore a neighbourhood of G can be determined (see § 113) so as to contain no points of κ; this neighbourhood can also be so determined as to be interior to K, and therefore to O. In this neighbourhood the condition (1), or (2), is satisfied.

---

* These lemmas were established by Bolza, see *Vorlesungen über Variationsprinzip*, pp. 155–158. The proofs in the text were given by Hobson, *Proc. Lond. Math. Soc.* (2), vol. XIV (1914), p. 151.

If $f(x_1, x_2, \dots x_n)$ is continuous in an open set O, and if it have the value zero at all points of a closed set G, interior to O, a neighbourhood of G can be determined, interior to O, such that at every point of that neighbourhood

$$|f(x_1, x_2, \dots x_n)|$$

is less than an assigned positive number ϵ.

Let a closed neighbourhood H, of G, be determined, that is contained in O. The points of H (if any) at which $|f| \geq \epsilon$ form a closed set L, which has no points in common with G. Determine a neighbourhood of G that contains no points of L, and is interior to H. At every point of this neighbourhood the condition $|f| < \epsilon$ is satisfied, and it is interior to H, and therefore to O.

318. Let us now suppose that x, in $F(x, y)$, denotes a point

$$(x_1, x_2, \dots x_m)$$

of an m-dimensional set, and further that $F(x, y)$ involves p parameters $c_1, c_2, \dots c_p$. We may then state the following generalization of the theorem of § 316.

Let $c_1, c_2, \dots c_p$ have the systems of values corresponding to points in a p-dimensional open domain X_p, and let the $m + 1$ variables $x_1, x_2, \dots x_m, y$ have the systems of values representing the points in an $(m + 1)$-dimensional open domain D_{m+1}; the $p + m + 1$ variables having accordingly the systems of values representing the points in an open domain D_{p+m+1}. Let

$$F(x_1, x_2, \dots x_m, y; c_1, c_2, \dots c_p)$$

be defined for all points in D_{p+m-1}, and be such as to satisfy the following conditions:

(1). *F is continuous, and $\dfrac{\partial F}{\partial y}$ exists, and is continuous, in D_{p+m+1}.*

(2). *At a fixed point $(\alpha_1, \alpha_2, \dots \alpha_m, \beta)$ in D_{m+1}, F has the value zero, and $\dfrac{\partial F}{\partial y} \neq 0$, for all points $(c_1, c_2, \dots c_p)$ in X_p.*

If δ be an arbitrarily chosen positive number sufficiently small, another such number ϵ can be determined, such that, to each value of y for which $|y - \beta| < \delta$, there corresponds one and only one set of values of

$$(x_1, x_2, \dots x_m),$$

for which $|x_r - \alpha_r| < \epsilon$, for $r = 1, 2, 3, \dots m$, and for each point of any given closed domain $\bar{X}_p$ interior to X_p; and these values of $x_1, x_2, \dots x_m, y$ are such that $F(x_1, x_2, \dots x_m, y; c_1, c_2, \dots c_p) = 0$. Thus a unique function

$$y = \phi(x_1, x_2, \dots x_m; c_1, c_2, \dots c_p)$$

is determined.

Moreover, if $\frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \dots \frac{\partial F}{\partial x_m}$, *and* $\frac{\partial F}{\partial c_1}, \frac{\partial F}{\partial c_2}, \dots \frac{\partial F}{\partial c_p}$ *all exist, and are continuous in* D_{p-m-1}, *the function* ϕ *has continuous partial differential coefficients with respect to the* $m+p$ *variables* $x_1, x_2, \dots x_m, c_1, c_2, \dots c_p$.

The proof is similar to that in § 316, but it is necessary to employ the lemmas in § 317.

The function $F(x_1, x_2, \dots x_m, y; c_1, c_2, \dots c_p)$ has the value zero, when

$$x_1 = \alpha_1, x_2 = \alpha_2, \dots x_m = \alpha_m, y = \beta,$$

whatever set of values $c_1, c_2, \dots c_p$ may have in X_p. Since $\frac{\partial F}{\partial y}$ is, for the same sets of values, always different from zero, and is continuous in X_p, it has everywhere the same sign in X_p; let it be assumed that that sign is positive. A neighbourhood (h) of the set of points

$$(\alpha_1, \alpha_2, \dots \alpha_m, \beta; c_1, c_2, \dots c_p),$$

where $c_1, c_2, \dots c_p$ is in $\bar{X}_p$, a closed set interior to X_p, can be so determined that, for all points in that neighbourhood, $\frac{\partial F}{\partial y}$ is positive. Thus for

$$\alpha_r - h < x_r < \alpha_r + h, (r = 1, 2, 3, \dots m), \quad \beta - h < y < \beta + h, (c_1, c_2, \dots c_p)$$

in $\bar{X}_p$, $\frac{\partial F}{\partial y}$ is positive. If $x_1, x_2, \dots x_m$ have the values $\alpha_1, \alpha_2, \dots \alpha_m$, since $\frac{\partial F}{\partial y} > 0$, as y is increased from $\beta - h$ to $\beta + h$, F is increased, and therefore, since it is zero when $y = \beta$, it must be negative when $\beta - h \leqq y < \beta$, and it must be positive when $\beta < y < \beta + h$. This holds for every point of $\bar{X}_p$.

A neighbourhood of the $(m+p)$-dimensional set of points

$$(\alpha_1, \alpha_2, \dots \alpha_m; c_1, c_2, \dots c_p)$$

at which, for $y = \beta - h$, F is negative, can be so determined that F is negative in that neighbourhood. Thus, for a properly chosen positive number k_1 $(\leqq h)$,

$$F(x_1, x_2, \dots x_m, \beta - h; c_1, c_2, \dots c_p) < 0$$

for $\quad \alpha_r - k_1 < x_r < \alpha_r + k_1, (r = 1, 2, 3, \dots m)$.

Similarly, a positive number k_2 $(\leqq h)$ can be so determined that

$$F(x_1, x_2, \dots x_m, \beta + h; c_1, c_2, \dots c_p) > 0$$

for $\quad \alpha_r - k_2 < x_r < \alpha_r + k_2, (r = 1, 2, 3, \dots m)$.

If k be the smaller of the two numbers k_1, k_2, we see that, for

$$\alpha_r - k < x_r < \alpha_r + k, (r = 1, 2, 3, \dots m),$$

F is negative when $y = \beta - h$, and positive when $y = \beta + h$; whatever point $(c_1, c_2, \dots c_p)$ may be of $\bar{X}_p$. Since $\frac{\partial F}{\partial y} > 0$, for all points such that

$$\alpha_r - k < x_r < \alpha_r + k, (r = 1, 2, \dots m), \beta - h < y < \beta + h, (c_1, c_2, \dots c_p) \text{ in } \bar{X}_p,$$

we see that, for each point of $\bar{X}_p$ there is one point y, and only one, such that $| y - \beta | < h$, at which F vanishes, for each set of values of

$$x_1, x_2, \dots x_m.$$

Thus a function $y = \phi(x_1, x_2, \dots x_m; c_1, c_2, \dots c_p)$ is determined, for

$$| x_r - a_r | < k, \ (r = 1, 2, \dots m),$$

and

$$(c_1, c_2, \dots c_p) \text{ in } \bar{X}_p,$$

such that F vanishes when y has the value of ϕ.

The difference of the values of F at two points

$$(x_1, x_2, \dots x_m, y; c_1, c_2, \dots c_p),$$

$$(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots x_m + \Delta x_m, y + \Delta y; c_1 + \Delta c_1, c_2 + \Delta c_2, \dots c_p + \Delta c_p)$$

is given by

$$\Delta F = \sum_{r=1}^{r=m} \left(\frac{\partial F}{\partial x_r} + \sigma_r \right) \Delta x_r + \left(\frac{\partial F}{\partial y} + \rho \right) \Delta y + \sum_{r=1}^{r=p} \left(\frac{\partial F}{\partial c_r} + \tau_r \right) \Delta c_r ;$$

where all the $m + p + 1$ letters σ_r, ρ, τ_r denote continuous functions that converge to zero, uniformly for all values of the variables in the closed set $\bar{D}_{m+p+1}$ interior to D_{m+p+1}.

When $$y = \phi(x_1, x_2, \dots x_m, c_1, c_2, \dots c_p),$$

and $$y + \Delta y = \phi(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots c_1 + \Delta c_1, c_2 + \Delta c_2, \dots),$$

we have $\Delta F = 0$. Thus if $\Delta c_1, \Delta c_2, \dots \Delta c_p$ all vanish, and

$$\Delta x_1, \Delta x_2, \dots \Delta x_{r-1}, \Delta x_{r+1}, \dots \Delta x_m$$

also vanish, we have

$$\frac{\Delta y}{\Delta x_r} = - \left(\frac{\partial F}{\partial x_r} + \sigma_r \right) \Big/ \left(\frac{\partial F}{\partial y} + \rho \right);$$

and since σ_r, ρ converge uniformly to zero, we have

$$\frac{\partial \phi}{\partial x_r} = - \frac{\partial F}{\partial x_r} \Big/ \frac{\partial F}{\partial y}.$$

In a similar manner, we see that

$$\frac{\partial \phi}{\partial c_r} = - \frac{\partial F}{\partial c_r} \Big/ \frac{\partial F}{\partial y}.$$

Thus the second part of the theorem has been established.

319. The theorem of § 318 may be extended to the case in which y is replaced by n variables $y_1, y_2, \dots y_n$, and there are now n functions $F_1, F_2, \dots F_n$ which involve the $m + n + p$ variables. The theorem then takes the following form:

Let $c_1, c_2, \dots c_p$ have the system of values corresponding to points in a p-dimensional open domain X_p, and let $(x_1, x_2, \dots x_m; y_1, y_2, \dots y_n)$ have the system of values corresponding to the points in an $(m + n)$-dimensional *open domain D_{m+n}; the $m + n + p$ variables having accordingly the values represented by points in an open domain D_{m+n+p}. Let n functions*

$$F_r(x_1, x_2, \dots x_m, y_1, y_2, \dots y_n; c_1, c_2, \dots c_p), \; (r = 1, 2, 3, \dots n)$$

be defined for all points in D_{m+n+p}, and be such as to satisfy the following conditions:

(1). The functions F_r are continuous, and all the functions

$$\frac{\partial F_r}{\partial y_s}, \; (s = 1, 2, \dots n),$$

exist and are continuous, in D_{m+n+p}.

(2). At a fixed point $(\alpha_1, \alpha_2, \dots \alpha_m, \beta_1, \beta_2, \dots \beta_n)$ in D_{m+n}, all the r functions vanish, and the Jacobian

$$\frac{\partial (F_1, F_2, \dots F_n)}{\partial (y_1, y_2, \dots y_n)}$$

has a value which is not zero, for all points $(c_1, \dots c_p)$ in X_p.

If δ be an arbitrarily chosen positive number, sufficiently small, another such number ϵ can be so determined that, to each set of values of $y_1, y_2, \dots y_n$ for which $| y_r - \beta_r | < \delta$, $(r = 1, 2, 3, \dots n)$, there corresponds one, and only one, set of values of $x_1, x_2, \dots x_m$ for which $| x_r - \alpha_r | < \epsilon$, $(r = 1, 2, \dots m)$, for each set of values of $(c_1, c_2, \dots c_p)$ in a closed domain $\overline{X}_p$, interior to X_p; and these values of x and y are such that the n functions F_r are all zero. Thus unique functions $y_r = \phi_r(x_1, x_2, \dots x_m; c_1, c_2, \dots c_p)$, $(r = 1, 2, 3, \dots n)$ are determined.

Moreover, if $\frac{\partial F_r}{\partial x_s}, \frac{\partial F_r}{\partial c_t}$ exist, for

$$r = 1, 2, \dots n; \; s = 1, 2, 3, \dots m; \; t = 1, 2, 3, \dots p,$$

and are continuous in D_{m+n+p}, the functions ϕ_r are all continuous and have continuous partial differential coefficients with respect to the $m + p$ variables.

The theorem of § 318 is the particular case of this theorem which arises when $n = 1$. Assume that the theorem holds when $n = \nu - 1$; it will then be shewn to hold when $n = \nu$, and thus it will hold generally.

As before, if
$$\frac{\partial (F_1, F_2, \dots F_\nu)}{\partial (y_1, y_2, \dots y_\nu)} > 0,$$

where $\quad x_r = \alpha_r, \; (r = 1, 2, \dots m), \quad y_r = \beta_r, \; (r = 1, 2, 3, \dots \nu),$

for all points $(a_1, a_2, \dots a_p)$ in X_p, there exists a positive number h, such that the same inequality holds, provided $| x_r - \alpha_r | < h$, $| y_r - \beta_r | < h$, for all points of $\overline{X}_p$ a closed domain contained in X_p.

At a point at which the Jacobian is not zero it is impossible that all the partial differential coefficients $\frac{\partial F_1}{\partial y_1}, \frac{\partial F_1}{\partial y_2}, \dots \frac{\partial F_1}{\partial y_\nu}$ can vanish. Choosing a number h' ($< h$), at each point of the closed domain E given by

$$|y_r - \beta_r| \leq h', (r = 1, 2, 3, \dots \nu),$$

$$|x_r - \alpha_r| \leq h', (r = 1, 2, \dots m), \text{ and } (c_1, c_2, \dots c_p)$$

in $\bar{X}_p$, one at least of the above partial differential coefficients is different from zero. For any point of the closed set E, a closed neighbourhood of the point can be determined, in the whole of which one and the same differential coefficient is different from zero. By employing the Heine-Borel theorem, we see that a finite number of the cells that constitute these neighbourhoods exists, such that each point of E is interior to one at least of the cells of this finite set. We obtain in this manner a division of the set E into a finite number of parts, such that, in each part, one of the differential coefficients $\frac{\partial F_1}{\partial y_1}, \frac{\partial F_1}{\partial y_2}, \dots \frac{\partial F_1}{\partial y_\nu}$ is different from zero in the whole of that part. Let us assume that, in one of these parts, $\frac{\partial F_1}{\partial y_1} > 0$, and that this part contains the points

$$(\alpha_1, \alpha_2, \dots \alpha_m, \beta_1, \beta_2, \dots \beta_\nu; c_1, c_2, \dots c_p)$$

in its interior, provided that $(c_1, c_2, \dots c_p)$ belongs to a certain closed part of $\bar{X}_p$ which we may denote by X_p^0.

We now consider $F_1(x_1, x_2, \dots x_m, y_1, y_2, \dots y_\nu; c_1, c_2, \dots c_p)$ as a function of $x_1, x_2, \dots x_m, y_1$, and of the parameters $y_2, y_3, \dots y_\nu, c_1, c_2, \dots c_p$. Applying now the theorem of § 317, we see that positive numbers h_1, k_1 ($\leq h'$) can be so determined that one and only one value of y_1 exists, such that $|y_1 - \beta_1| < h_1$ corresponding to each set of values of

$$x_1, x_2, \dots x_m$$

such that $|x_r - \alpha_r| < k_1$, ($r = 1, 2, 3, \dots m$), and provided

$$(y_2, y_3, \dots y_\nu, c_1, c_2, \dots c_p)$$

is in a certain closed domain of $\nu + p - 1$ dimensions. This value of y_1 may be denoted by

$$y_1 = f_1(x_1, x_2, \dots x_m, y_2, y_3, \dots y_\nu; c_1, c_2, \dots c_p).$$

On substitution of this value of y_1 in the equations

$$F_2 = 0, F_3 = 0, \dots F_\nu = 0,$$

we obtain a set of $\nu - 1$ equations which may be denoted by

$$\phi_r(x_1, x_2, \dots x_m, y_2, y_3, \dots y_\nu; c_1, c_2, \dots c_p) = 0,$$

where $r = 2, 3, \dots \nu$. These hold when $|x_r - \alpha_r| < k_1$, and provided

$$(y_2, y_3, \dots y_\nu; c_1, c_2, \dots c_p)$$

is in the specified domain. Assuming the theorem to hold for $n = \nu - 1$, we see that there exist values of $y_2, y_3, \dots y_\nu$, given by equations

$$y_r = \phi_r (x_1, x_2, \dots x_m : c_1, c_2, \dots c_p) \; (r = 2, 3, \dots \nu),$$

and that these values are unique for every set of values of

$$x_1, x_2, \dots x_m, c_1, c_2, \dots c_p, \text{ such that } | x_r - \alpha_r | < k_2 \; (r = 1, 2, \dots m),$$

and such that $(c_1, c_2, \dots c_p)$ is in $X_p{}^0$; where k_2 is some positive number. The values of $y_2, y_3, \dots y_\nu$ are such that $| y_r - \beta_r | < h_3$, some fixed positive number: this must be so small that the points $(y_2, y_3, \dots y_\nu, c_1, c_2, \dots c_p)$ are all in the specified closed domain. This holds, subject to the condition that

$$\frac{\partial (\phi_2, \phi_3, \dots \phi_\nu)}{\partial (y_2, y_3, \dots y_\nu)} \neq 0$$

at the points $(\alpha_1, \alpha_2, \dots \alpha_m, \beta_2, \beta_3, \dots \beta_\nu; c_1, c_2, \dots c_p)$,

where $(c_1, c_2, \dots c_p)$ is in $X_p{}^0$.

The function f_1 has continuous partial differential coefficients with respect to all the variables contained in it. Thus

$$\frac{\partial f_1}{\partial y_r} = - \frac{\partial F_1}{\partial y_r} \Big/ \frac{\partial F_1}{\partial y_1}, \; (r = 2, 3, \dots \nu)$$

$$\frac{\partial f_1}{\partial x_s} = - \frac{\partial F_1}{\partial x_s} \Big/ \frac{\partial F_1}{\partial y_1}, \; (s = 1, 2, \dots m)$$

$$\frac{\partial f_1}{\partial c_q} = - \frac{\partial F_1}{\partial c_q} \Big/ \frac{\partial F_1}{\partial y_1}, \; (q = 1, 2, \dots p).$$

Now $$\frac{\partial \phi_r}{\partial y_s} = \frac{\partial F_r}{\partial y_s} + \frac{\partial F_r}{\partial y_1} \left(- \frac{\partial F_1}{\partial y_s} \Big/ \frac{\partial F_1}{\partial y_1} \right).$$

On substituting the values of the partial differential coefficients in the Jacobian, we obtain, after a slight transformation,

$$\frac{\partial (F_1, F_2, \dots F_\nu)}{\partial (y_1, y_2, \dots y_\nu)} \Big/ \frac{\partial F_1}{\partial y_1};$$

and thus the condition is that

$$\frac{\partial (F_1, F_2, \dots F_\nu)}{\partial (y_1, y_2, \dots y_\nu)} \neq 0$$

at the points $(\alpha_1, \alpha_2, \dots \alpha_m, \beta_1, \beta_2, \dots \beta_\nu; c_1, c_2, \dots c_p)$, where $(c_1, c_2, \dots c_p)$ is in $X_p{}^0$.

It has now been shewn that there is a unique value of y_1, such that $| y_1 - \beta_1 | < h_1$, for each set of values of

$$y_2, y_3, \dots y_\nu, x_1, x_2, \dots x_m, c_1, c_2, \dots c_p,$$

such that $(y_2, y_3, \dots y_\nu, c_1, c_2, \dots c_p)$ is in a certain closed domain which includes a set of points for which

$$y_2 = \beta_2, y_3 = \beta_3, \dots y_\nu = \beta_\nu; \text{ and that } | x_r - \alpha_r | < k_1.$$

It has also been shewn that a unique set of values of $y_2, y_3, \dots y_\nu$, such that $| y_r - \beta_r | < h_2$, $(r = 2, 3, \dots \nu)$ can then be determined for each set of values of $(x_1, x_2, \dots x_m, c_1, c_2, \dots c_p)$ such that $|x_r - a_r| < h_2$ and that $(c_1, c_2, \dots c_p)$ is in X_p^0. We can choose any value of h_1, or of h_2, smaller than the value fixed: then k_1 and k_2 will in general have to be diminished. By altering the value of one of the numbers h_1, h_2 we may make them have equal values, say $\bar h$: and for $\bar k$ we then take the smaller of the corresponding values k_1, k_2. It now appears that there exists a unique set of values of $y_1, y_2, \dots y_\nu$, such that $|y_r - \beta_r| < \bar k$, $r = 1, 2, \dots \nu$, corresponding to each point

$$(x_1, x_2, \dots x_m, c_1, c_2, \dots c_p)$$

such that $|x_r - a_r| < \bar h$, $(r = 1, 2, \dots m)$, and that $(c_1, c_2, \dots c_p)$ is in X_p^0.

As every point of $\bar X_p$ is in one or more of a finite number of such sets as X_p^0, corresponding to each of which one of the partial differential coefficients $\frac{\partial F_1}{\partial y_r}$ is different from zero for the whole of that set, the above reasoning may be applied to each such set X_p^0.

We have, in each case, a pair of positive numbers $\bar h$, $\bar k$, determined as above; moreover we may take, instead of $\bar h$ any number less than $\bar h$; $\bar k$ being correspondingly diminished. If we take for δ the smallest of the numbers $\bar h$, it is clear that a corresponding number ϵ may be determined so that δ and ϵ can be used, instead of $\bar h$, $\bar k$, for all the parts of $\bar X_p$. It has thus been shewn that the theorem holds in the case $n = \nu$, if it is assumed to hold in the case $n = \nu - 1$. The theorem was proved in § 318, for the case $n = 1$; and therefore it holds generally.

It is easy to see that the above general theorem may be modified so as to apply to the establishment of unique values of $y_1, y_2, \dots y_n$ which satisfy the n equations

$$F_r(x_1, x_2, \dots x_m, y_1, y_2, \dots y_n; c_1, c_2, \dots c_p) = f_r(c_1, c_2, \dots c_p),$$

where the functions $f_r(c_1, c_2, \dots c_p)$ satisfy the conditions that they have continuous partial differential coefficients in the domain X_p: the functions F_r are to satisfy the same conditions as before.

We have, in fact, only to consider the functions $F_r - f_r$, instead of the functions F_r. It is then assumed, as before, that $F_r - f_r = 0$, when

$$x_r = a_r,\ y_r = \beta_r, \text{ for all values of } (c_1, c_2, \dots c_p) \text{ in } X_p.$$

An important case of the general theorem is that in which $n = m$, and in which F_r is of the form $\phi_r(y_1, y_2, \dots y_n; c_1, c_2, \dots c_p) - x_r$. We have then an extension of the theorem of inversion.

The case of the general theorem which arises when $c_1, c_2, \dots c_p$ are absent is the well-known theorem of Dini relating to implicit functions*.

---

* See Jordan's *Cours d'Analyse*, vol. I, §§ 91, 92; also Osgood's *Lehrbuch der Funktionentheorie*, vol I, pp. 47–57. On the general theory of functions defined implicitly, see W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. VII (1909), p. 397; Hobson, *Proc. Lond. Math. Soc.* (2), vol. XIV (1914), p. 147, and Hedrick and Westfall, *Bull de la soc. math. de France*, vol. XLVI (1916), p. 1.

MAXIMA AND MINIMA OF A FUNCTION OF TWO VARIABLES

320. Let us suppose that a function $f(x, y)$ is defined at all points in a two-dimensional neighbourhood of the point (x_0, y_0).

If the function be such that $f(x_0 + h, y_0 + k) - f(x_0, y_0) < 0$, for all values of h, k which are not both zero, and are such that $|h|$, $|k|$ are both less than some fixed positive number δ, then the function $f(x, y)$ is said to have *a proper maximum* at the point (x_0, y_0).

In case the fixed number δ can only be so determined that the condition $f(x_0 - h, y_0 + k) - f(x_0, y_0) \leqq 0$ is satisfied, the function is said to have *an improper maximum* at the point (x_0, y_0).

If the conditions contained in these definitions be replaced by

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) > 0, \text{ and } f(x_0 + h, y_0 + k) - f(x_0, y_0) \geqq 0$$

respectively, the function $f(x, y)$ is said to have, in the first case, *a proper minimum*, and in the second case, *an improper minimum*, at the point (x_0, y_0).

A proper or improper maximum or minimum may be spoken of as an *extreme* of the function.

At an extreme (x_0, y_0),

$$f(x_0 + h, y_0) - f(x_0, y_0), \quad f(x_0 - h, y_0) - f(x_0, y_0)$$

both have the same sign, or are zero, for all sufficiently small values of h; it follows that, if $\dfrac{\partial f(x_0, y_0)}{\partial x_0}$ exist, it must be zero. A similar remark applies to $\dfrac{\partial f(x_0, y_0)}{\partial y_0}$.

These conditions are necessary, under the hypothesis of the existence of the two partial differential coefficients, but not sufficient, for the existence of an extreme at the point (x_0, y_0).

If we write $x = x_0 - r\cos\theta$, $y = y_0 + r\sin\theta$, $f(x, y) = \phi(r, \theta)$, it is clearly necessary for the existence of an extreme of $f(x, y)$ at (x_0, y_0), that $\phi(r, \theta)$, for each constant value of θ, should have an extreme at $r = 0$. Thus, for an assigned value of θ, a positive number α_θ can be determined, such that one of the four conditions

$$\phi(r, \theta) - f(x_0, y_0) < 0, \quad \phi(r, \theta) - f(x_0, y_0) \leqq 0,$$
$$\phi(r, \theta) - f(x_0, y_0) > 0, \quad \phi(r, \theta) - f(x_0, y_0) \geqq 0,$$

according as the point is a proper maximum, an improper maximum, a proper minimum, or an improper minimum, shall be satisfied for all values of r, different from zero, and such that $|r| < \alpha_\theta$. Thus an extreme of a function is necessarily an extreme for values of the function on each straight line drawn through the point.

This condition, though necessary, is however not sufficient; for a_θ may have a definite value for each value of θ, and yet the lower boundary of a_θ, for all values of θ, may be zero. In this case, no value of δ can be determined, as required in the definition of the extreme in the two-dimensional domain. It has thus been shewn that, *in order that (x_0, y_0) may be an extreme point for the function $f(x, y)$, it is necessary and sufficient*[*] *(1) that $r = 0$ should be an extreme point of $\phi(r, \theta)$ for each value of θ, and (2) that the number a_θ which is so determined for each value of θ that, for $r < a_\theta$, the condition as to $\phi(r, \theta) - f(x_0, y_0)$ may be satisfied, should have a finite lower boundary, when all values of θ, $(0 \leq \theta < \pi)$ are considered. If the lower boundary of a_θ be zero, the point is not an extreme point of the function.*

When the lower boundary of a_θ is $d\ (> 0)$, the neighbourhood of (x_0, y_0), which must exist in accordance with the definition, is the square of which the corners are the four points $\left(x_0 \pm \frac{d}{\sqrt{2}}, y_0 \pm \frac{d}{\sqrt{2}}\right)$.

EXAMPLE

As an example of a function which possesses no minimum at a point, although the point is a minimum for each straight line through the point, we may take the function

$$(y - ax^2)(y - bx^2) \equiv y^2 - y(ax^2 + bx^2) - abx^4,$$

where a and b have positive values.

The function is positive outside the two parabolas $y - ax^2 = 0,\ y - bx^2 = 0$, and in the space interior to the inner parabola; in the space between the parabolas, the function is negative. Along any straight line QAR through A (0, 0), the function exceeds $f(0, 0)$ at all points interior to AP, and everywhere in PA produced; thus for the line QAR the function has a minimum at A. The point (0, 0) is not a minimum of the function, since the lower limit of AP for all positions of QAR is zero; and thus there exists no two-dimensional neighbourhood of A, in which the function is never less than at A.

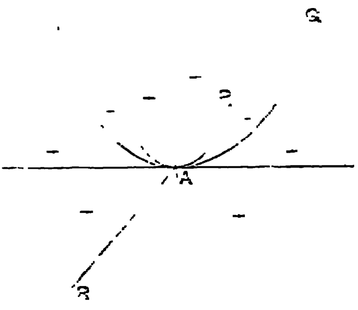

FIG. 4

321. We may, without loss of generality, take the point at which the conditions for the existence of an extreme of the function $f(x, y)$ are to be investigated to be the point (0, 0). It will be assumed that, at all points in the neighbourhood of (0, 0), $f(x, y)$ is continuous with respect to x, and also with respect to y. The following theorem contains a criterion for the existence of a proper maximum (minimum) at the point (0, 0).

---

[*] The necessity for this condition has been disregarded in many text-books. The insufficiency of (1) was first pointed out by Peano, *Calcolo diff.*, Turin, 1884, p. 29, in connection with the example given in the text. See also Dantscher, *Math. Annalen*, vol. XLII (1893), p. 89, and Scheeffer, *Math. Annalen*, vol. XXXV (1890), p. 541.

The necessary and sufficient conditions that the point (0, 0) may be a point at which $f(x, y)$ has a proper maximum (minimum) are the following*:—(1). *A positive number δ must exist which is such that, if x be any number different from zero, and numerically less than δ, the upper (lower) limit of $f(x, y)$, for such constant value of x, and for all values of y for which $-x \leq y \leq x$, being $f(x, \phi(x))$, this upper (lower) limit is, for every value of x $(-\delta < x \neq 0 < \delta)$, less (greater) than $f(0, 0)$.*

(2). *A positive number δ' must exist which is such that, if y be any number different from zero, and numerically less than δ', the upper (lower) limit of $f(x, y)$, for such constant value of y, and for all values of x for which*

$$-y \leq x \leq y,$$

being $f(\psi(y), y)$, this upper (lower) limit is, for every value of

$$y\ (-\delta' < y \neq 0 < \delta'),$$

less (greater) than $f(0, 0)$.

It will be observed that, since $f(x, y)$ is assumed to be continuous with respect to x, and also with respect to y, the limit $f(x, \phi(x))$ is actually attained for some value $\phi(x)$, of y, in the interval $(-x, x)$, and the limit $f(\psi(y), y)$ is actually attained for some value $\psi(y)$, of x, in the interval $(-y, y)$. It is clear that, unless both the conditions stated in the theorem be satisfied, $f(0, 0)$ cannot be a proper maximum (minimum) of the function. If, for example, no such number as δ in (1) can be determined, there are points in every neighbourhood of (0, 0) at which $f(x, y)$ is $\geq (\leq) f(0, 0)$.

The conditions are sufficient. For, if δ, δ' exist, the value of $f(x, y)$ at every point, except (0, 0), within the neighbourhood, the corners of which are the four points $(\pm\delta'', \pm\delta'')$, is less (greater) than $f(0, 0)$, where δ'' is the lesser of the two numbers δ, δ'.

The necessary and sufficient conditions that the function $f(x, y)$ may have an improper maximum (minimum) at (0, 0) are similar to the above. In this case $f(x, \phi(x))$ must be less than, or equal to (greater than, or equal to) $f(0, 0)$, for all the values of x in the interval, and $f(\psi(y), y)$ must be less than, or equal to (greater than, or equal to) $f(0, 0)$, for all values of y in the interval. Further, corresponding to every positive number $\bar{\delta} < \delta$, there must be a value of x $(< \bar{\delta})$, for which

$$f(x, \phi(x)) = f(0, 0);$$

or else a similar condition must hold for $f(\psi(y), y)$; or in both cases, the condition may be satisfied.

Other methods of determining whether (0, 0) be a point at which there is a maximum or minimum of $f(x, y)$ will be dealt with in Vol. II.

---

* See Stolz, *Wiener Sitzungsber.*, vol. CII *a* (1891), p. 1167; also *Grundzüge*, vol. I (1893), p. 213.

PROPERTIES OF A FUNCTION CONTINUOUS WITH RESPECT TO EACH VARIABLE

322. Let a function $f(x, y)$, defined for all values of x and y in a continuous domain, be everywhere continuous with respect to y, and be also continuous with respect to x for every straight line parallel to the x-axis, belonging to a set cutting the y-axis in an everywhere dense set of points.

Let A be the point (x, y), and let BC be drawn with A as its middle point, parallel to the y-axis, and of length 2ρ. If $\omega(\rho)$ be the fluctuation of $f(x, y)$ in the interval BC, then $\omega(\rho)$ is a continuous function of ρ; and $\lim\limits_{\rho \sim 0} \omega(\rho) = 0$, since $f(x, y)$ is everywhere continuous with respect to y. Let σ be a fixed positive number, and let $\beta_\sigma(x, y)$ denote the upper limit of those values of ρ for which $\omega(\rho) \leq \sigma$; thus $\omega(\rho) \leq \sigma$, if $\rho \leq \beta_\sigma(x, y)$; and $\omega(\rho) > \sigma$, if $\rho > \beta_\sigma(x, y)$.

The function $\beta_\sigma(x, y)$, thus defined for every point (x, y), is everywhere positive; and it will be shewn to be an upper semi-continuous function with respect to the two-dimensional domain (x, y), in accordance with the definition in § 230 and § 234.

Take $B_0A_0 = A_0C_0 = \beta_\sigma(x_0, y_0)$, and also $B_1B_0 = C_0C_1 = \epsilon$, where ϵ is a fixed positive number. The fluctuation of $f(x, y)$ in B_1C_1 is greater than σ; let it be $\sigma + k$. If k_1 be a fixed positive number $< k$, two points M, N can be found in B_1C_1, such that

$$|f(M) - f(N)| > \sigma + k_1.$$

Moreover, these points M, N can be so chosen as to lie on two straight lines parallel to the x-axis, which belong to the set along each of which $f(x, y)$ is continuous with respect to x; this follows from the fact that this set of straight lines cuts B_1C_1 in an everywhere dense set of points. Since $f(x, y)$ is continuous with respect to x, at each of the points M, N, two segments $M'M''$, $N'N''$, with M and N as their middle points, can be determined, so as to have equal lengths 2δ, and to be such that

$$|f(P) - f(M)| < \tfrac{1}{2}k_1,$$
$$|f(Q) - f(N)| < \tfrac{1}{2}k_1,$$

provided P be any point in $M'M''$, and Q be any point in $N'N''$.

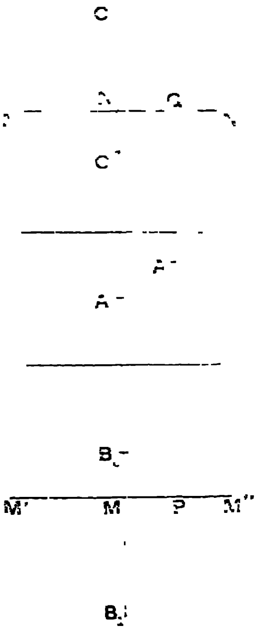

FIG. 5

From these inequalities and the former one, we deduce that

$$|f(P) - f(Q)| > \sigma.$$

Take the square of which A_0 is the centre, and of which the sides are parallel to the axes, and are at a distance from A_0 less than the smaller of the two numbers $\frac{1}{2}\epsilon$ and δ. If A be any point in this square, the distance of A from each of the straight lines $M'M''$, $N'N''$ is less than $\beta_\sigma(x_0, y_0) - \epsilon$. Through A let a straight line be drawn parallel to the y-axis, and mark off on it the segment of which A is the centre, and of which the half-length is $\beta_\sigma(x_0, y_0) - \epsilon$; this segment will cut $M'M''$ and $N'N''$, and therefore contains two points P, Q which are such that

$$|f(P) - f(Q)| > \sigma.$$

Therefore the fluctuation of $f(x, y)$ in this segment is $> \sigma$, and hence, at the point A (x, y), we have $\beta_\sigma(x, y) < \beta_\sigma(x_0, y_0) + \epsilon$. A square having been determined with its centre at A_0, such that for every point in this square

$$\beta_\sigma(x, y) < \beta_\sigma(x_0, y_0) + \epsilon,$$

it follows that $\beta_\sigma(x, y)$ is an upper semi-continuous function at A_0, with respect to the two-dimensional domain (x, y).

323. Let us now consider the linear set C, of points (x, y) defined by $y = \phi(x)$, where $\phi(x)$ is a continuous function of x. At each point of C, the function $\beta_\sigma(x, y)$ is defined, and it has at every point a minimum relatively to the set C, the term being used in accordance with the definition given in § 224.

If, at a point $A_0(x_0, y_0)$, of C, the function $\beta_\sigma(x, y)$ have its minimum with respect to C positive, then we shall prove that the saltus of $f(x, y)$ at A_0, with respect to the two-dimensional domain (x, y), is $\leqq 2\sigma$. Let γ denote this minimum, and let γ_1 be a positive number $< \gamma$.

Let an interval $(x_0 - \delta, x_0 - \delta)$ on the line $y = y_0$ be so determined that in this interval

$$|\phi(x) - \phi(x_0)| < \tfrac{1}{2}\gamma_1.$$

This interval may, if necessary, be so reduced, that for all values of x in it,

$$\beta_\sigma\{x, \phi(x)\} > \gamma_1.$$

Describe the rectangle R, with A_0 as centre, the sides parallel to the axes of x and y being 2δ and γ_1 respectively. On every segment PQ, of R, parallel to Oy, the fluctuation of the function is $\leqq \sigma$; for there is on PQ a point A of the set C, and the segment with centre A, and length $2\beta_\sigma(A) > 2\gamma_1$, contains the whole segment PQ.

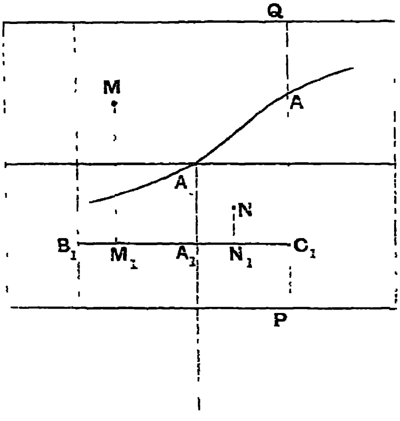

FIG. 6

Taking a fixed positive number ϵ, an area surrounding A_0 can be determined, in which the fluctuation of $f(x, y)$ is $\leqq 2\sigma + \epsilon$. To effect this, take a point A_1 on the ordinate through A_0, and in the rectangle R, such that A_1 is a point of continuity of $f(x, y)$ with respect to x; then on the straight line $y = y_1$, take a segment B_1C_1, with centre A_1, and of length $2\delta' < 2\delta$, such that the fluctuation of f in B_1C_1 is $< \epsilon$. Consider the rectangle R' contained in R, such that the sides of R' are of lengths $2\delta'$ and 2η, parallel to the axes, its centre being at A_1. The fluctuation of $f(x, y)$ in this rectangle is $< 2\sigma + \epsilon$. For if M, N be any two points in it, let M_1, N_1 be their projections on B_1C_1; then

$$|f(M) - f(M_1)| \leq \sigma, \quad |f(N) - f(N_1)| \leq \sigma, \quad |f(M_1) - f(N_1)| < \epsilon;$$

from these inequalities we deduce that

$$|f(M) - f(N)| < 2\sigma + \epsilon.$$

Since this holds for every ϵ, the saltus of $f(x, y)$ at A_1 is $\leq 2\sigma$. If, at a point A_0, the saltus of $f(x, y)$ be $> 2\sigma$, then at A_0 the minimum of β_σ with respect to C must be zero.

Since β_σ is positive at every point of C, and is an upper semi-continuous function of (x, y), it follows from the theorem of § 232, that in every arc of the curve C, there exists an arc D_1 in which the minimum of β_σ is positive. Let us take a sequence $\sigma_1, \sigma_2, \ldots \sigma_n, \ldots$ of positive decreasing numbers, of which the limit is zero. It is then clear that in every arc D, there exists a point where β_{σ_n} has its minimum, with respect to C, positive, for every σ_n. At this point the fluctuation of $f(x, y)$ with respect to the two-dimensional domain (x, y) is $\leq 2\sigma_n$ for all values of n, and is therefore zero. This point must be a point of continuity of $f(x, y)$ with respect to (x, y).

The following general theorem* has thus been established:

If $f(x, y)$ be a function of the two variables x, y, which is everywhere continuous with respect to y, and is continuous with respect to x along straight lines parallel to the x-axis, which cut the y-axis in an everywhere dense set of points, then in every portion of a curve $y = \phi(x)$, where $\phi(x)$ is a continuous function, there exist points at which $f(x, y)$ is continuous with respect to the two-dimensional domain (x, y).

It follows from this theorem that points of continuity exist in every area, that is $f(x, y)$ *is at most a point-wise discontinuous function*.

The whole of the reasoning above is applicable, if only those points of (x, y) are taken into account which belong to a perfect set G. It thus appears that, under the conditions stated in the above theorem, $f(x, y)$ is a point-wise discontinuous function relatively to every perfect set G, of points in (x, y). The points of continuity of $f(x, y)$, on the curve $y = \phi(x)$, are everywhere dense with respect to every perfect set of points on the curve.

---

* Baire, *Annali di Mat.* Ser. III^a, vol. III (1899), p. 27.

EXAMPLES

1. If* $f(x, y)$ be a function which is everywhere continuous with respect to each of the variables x, y, then the points at which the saltus of $f(x, y)$, with respect to the two-dimensional continuum (x, y), is $\geqq \sigma$, form a set of points such that the projection of the set on either axis, by lines parallel to the other axis, is a non-dense set.

2. If† a function $f(x, y, z)$ of three variables x, y, z be everywhere continuous with respect to each variable, then $f(x, y, z)$ is at most a point-wise discontinuous function relatively to the three-dimensional continuum (x, y, z). Further, on every surface $x=\phi(y, z)$, where ϕ is continuous with respect to (y, z), the function $f(x, y, z)$ is at most a point-wise discontinuous function with respect to (y, z). The set of points at which the saltus of $(x, y, z) \geqq \sigma$ may contain all the points of a continuous curve

3. Let† $\phi(x, y)$ be a function which is continuous with respect to each of the variables x and y, and let $(0, 0)$ be a point of discontinuity of $\phi(x, y)$ with respect to (x, y). Define $f(x, y, z)$ by the condition $f(x, y, z)=\phi(x, y)$; then the function $f(x, y, z)$ is continuous with respect to each of the three variables, but every point on the z-axis is a point of discontinuity with respect to (x, y, z).

4. Let† $f(x, y, z)$ be a function which is constant along any straight line parallel to the straight line $x=y=z$, and is such that $f(x, y, 0)=\frac{xy(x-y)}{(x^2+y^2)^{\frac{3}{2}}}$, $f(0, 0, 0)=0$. This function is discontinuous at every point on the straight line $x=y=z$.

5. Let‡ $f(0, 0)=0$; at all points at which y is positive and $x^2/y \leq 1$, let $f(x, y)=x^2/y$, and when $x^2/y \geqq 1$, let $f(x, y)=y/x^2$. Also let $f(x, -y)=f(x, y)$. The origin is a point of discontinuity, at which the saltus is 1; elsewhere the function is continuous with respect to (x, y). The function is however continuous with respect to every straight line: for it is continuous along the axis of x, where it has the value 0. Consider now the straight line $y=mx$, where $m \neq 0$; it lies entirely in the part of the plane in which

$$|f(x, y)|=x^2/|y|=|x/m|,$$

so that, as the point $(0, 0)$ is approached, $f(x, y)$ converges to 0.

6. Let‡ $(a_1, b_1), (a_2, b_2), \ldots (a_n, b_n) \ldots$ denote an enumerable set of points, and let

$$F(x, y)=\frac{1}{2}f(x-a_1, y-b_1)+\frac{1}{2^2}f(x-a_2, y-b_2)+\ldots+\frac{1}{2^n}f(x-a_n, y-b_n)+\ldots,$$

where $f(x, y)$ denotes the function defined in Ex. 5.

The function $F(x, y)$ is continuous with respect to any straight line, and is continuous with respect to (x, y) at all points except those of the enumerable set. The enumerable set may be everywhere dense in the plane.

324. The methods developed by Baire of dealing with functions of two or more variables, in relation to the distribution of the points of discontinuity, have been applied by him to the consideration of the following three problems:

(1). What must be the nature of a function $\phi(x)$, defined for $\alpha \leqq x \leqq \beta$, in order that a function $f(x, y)$ can exist which is defined for all points in the square $\alpha \leq x \leq \beta$, $\alpha \leq y \leq \beta$, and is continuous at every point with respect to x and with respect to y, and moreover is equal to $\phi(x)$ on the straight line $x = y$?

---

* Baire, *loc. cit.* p. 94.

† Baire, *loc. cit.* p. 99.

‡ W. H. and G. C. Young, *Quarterly Journal of Math.* vol. XLI (1910), p. 87, where an example is also constructed of a function that is continuous with respect to every straight line but is discontinuous with respect to (x, y) at points of an everywhere dense unenumerable set.

(2). What must be the nature of a function $\phi(x)$, defined for $\alpha \leq x \leq \beta$ in order that a function $f(x, y)$ can be defined for all points in the square $\alpha \leq x \leq \beta$, $\alpha \leq y \leq \beta$, which shall satisfy the conditions that it is continuous with respect to (x, y) at every point for which $y > 0$, is continuous with respect to y at the points of $y = 0$, and is equal to $\phi(x)$ when $y = 0$?

(3). A function $f(x, y)$ is defined in the rectangle $\alpha \leq x \leq \beta$, $\gamma \leq y \leq \delta$, and is everywhere continuous with respect to y. Further, there is a set of parallels to the x-axis, along each of which $f(x, y)$ is continuous with respect to x; these parallels intersecting the straight line $x = \alpha$ in a set of points which is everywhere dense in the interval (γ, δ). What is the nature of the function $f(x, y)$ on a continuous curve drawn in the rectangle?

The problems (1), (2) are particular cases of (3). It has been shewn above that a necessary condition satisfied by $f(x, y)$, in (3), is that it should be a point-wise discontinuous function relatively to every perfect set of points. That this condition is also sufficient, has been demonstrated by Baire in his memoir quoted above. A proof of this will be given, for the case of problem (2), in Vol. II, in connection with the theory of functions representable as the limits of sequences of functions.

THE REPRESENTATION OF A SQUARE ON A LINEAR INTERVAL

325. Let a point of a square whose side is unity be denoted by (x, y), where $0 \leq x \leq 1$, $0 \leq y \leq 1$; and let t denote a point of a linear interval (0, 1). An account has been given in § 62 of Cantor's method of establishing a (1, 1) correspondence between the points of the square and those of the linear interval. Such a correspondence denotes functional relations $x = f(t)$, $y = \phi(t)$ between x, y as dependent variables, and t as an independent variable. It will be shewn however that no (1, 1) relation between the two sets of points can be a continuous representation*; *i.e.* it is impossible that the functions $f(t)$, $\phi(t)$ can be both continuous.

Let us assume that such a continuous representation can be defined. To any closed set of points $\{t\}$, in (0, 1), there will correspond a closed set in the plane area. For if $t_1, t_2, \ldots t_n, \ldots$ be a convergent sequence of points t, of which t_ω is the limiting point, then the point $f(t_\omega)$, $\phi(t_\omega)$ is the limiting point of the set of points (x_1, y_1), (x_2, y_2), … (x_n, y_n), … which correspond to $t_1, t_2, \ldots t_n, \ldots$ respectively; therefore to a closed set $\{t\}$

---

* See Netto, "Beitrag zur Mannigfaltigkeitslehre," *Crelle's Journal*, vol. LXXXVI (1879), p. 263, also Loria, *Giorn. di Mat.* vol. XXV (1887), p. 97. In the proof given by these writers it is assumed that a closed curve corresponds to a linear sub-interval of (0, 1); this is not necessarily the case, for a non-dense closed set may correspond to the closed curve.

there corresponds a closed set $\{(x, y)\}$. Again, to a convergent sequence (x_1, y_1), (x_2, y_2), ... of points in the plane area, there corresponds a set of points $t_1, t_2, \ldots$ in the linear interval, the latter of which has a limiting point t_ω which must correspond to (x_ω, y_ω); and since only one value of t corresponds to one set of values of (x, y), there can be only one such limiting point t_ω. Thus, to a closed set in the plane, there corresponds a closed set in the linear interval. Take two points t_1, t_2 in the interval $(0, 1)$: these points correspond to two points P_1, P_2 in the square area. To the closed linear interval (t_1, t_2) there corresponds a closed set S which contains the points P_1, P_2. It can be shewn that there are points other than P_1, P_2 on the frontier of S. Denote by $C(S)$ the set of those points of the square area which do not belong to S. Two points Q, R in the square can be determined, such that Q lies on the straight line P_1P_2, and R does not lie on this straight line; such that neither Q nor R coincides with P_1 or P_2, and such that one of the two belongs to S and the other to $C(S)$. The closed set consisting of the straight line QR contains points both of S and of $C(S)$: those points of S which lie on it form a closed set, and there must be one such point of S at least which is on the frontier of S; such a point may, or may not, coincide with Q or R. Since then S contains points on its frontier besides P_1 and P_2, we can take a point $\bar{t}$, within the linear interval (t_1, t_2), such that the point T in the square which corresponds to it is on the frontier of S. Since T is the limiting point of a sequence of points of $C(S)$, it follows that $\bar{t}$ must be the limiting point of a sequence of points all of which are external to the interval (t_1, t_2), and this is impossible. It has thus been established that:

No continuous (1, 1) *correspondence can exist between all the points in a square and all the points in a linear interval.*

In particular, the correspondence shewn by Cantor to exist must be discontinuous.

326. The reasoning of § 325 would be inapplicable if the correspondence $x = f(t)$, $y = \phi(t)$ were such that, to a given point (x, y) more than one point t may correspond, the functions $f(t)$, $\phi(t)$ being still one-valued continuous functions, so that if t be assigned, (x, y) is uniquely determined. In this case, the limiting point of the set of points external to the interval (t_1, t_2) would be not $\bar{t}$, but another value of t which also corresponds to the point T.

Peano* gave the first continuous correspondence of the kind just indicated, thus defining what, by a considerable extension of the use of geometrical language, may be called a continuous curve which passes through every point of the square at least once.

---

* "Sur une courbe, qui remplit toute une aire plane," *Math. Ann.* vol. XXXVI (1890), p. 157.

Let the points in the interval (0, 1) be expressed in the form

$$t = \cdot a_1 a_2 a_3 \ldots a_n \ldots$$

in radix fractions in the ternary scale, so that each a is either 0, 1, or 2. Let $k(a)$ denote the number $2 - a$, so that $k(2) = 0$, $k(1) = 1$, $k(0) = 2$; and let $k^n(a)$ denote the result of performing this operation n times, so that $k^n(a)$ is a or $2 - a$, according as n is even or odd.

Let x, y be defined, for a prescribed t, by

$$x = \cdot b_1 b_2 b_3 \ldots, \quad y = \cdot c_1 c_2 c_3 \ldots,$$

the ternary scale being again employed; the numbers b, c being defined by the relations

$$b_1 = a_1, \quad b_2 = k^{a_2}(a_3), \ldots b_n = k^{a_2 + a_4 + \ldots + a_{2n-2}}(a_{2n-1}),$$
$$c_1 = k^{a_1}(a_2), \quad c_2 = k^{a_1 + a_3}(a_4), \ldots c_n = k^{a_1 + a_3 + \ldots + a_{2n-1}}(a_{2n});$$

thus b_n is equal to a_{2n-1} or to $2 - a_{2n-1}$, according as $a_2 + a_4 + \ldots + a_{2n-2}$ is even or odd.

The numbers t may be divided into two classes:

(1). Those, other than 0 or 1, which are capable of a double representation

$$t = \cdot a_1 a_2 a_3 \ldots a_n 222 \ldots = \cdot a_1 a_2 \ldots \overline{a_n + 1}\, 000 \ldots.$$

(2). Those which have a single representation only.

If t be a number of the second class, x and y are uniquely defined. If t be a number of the first class

$$t = \cdot a_1 a_2 \ldots a_n 222 \ldots \equiv \cdot a_1 a_2 \ldots \overline{a_n + 1}\, 000 \ldots.$$

let $\cdot b_1 b_2 b_3 \ldots$, $\cdot b_1' b_2' b_3' \ldots$ denote the numbers obtained by applying the definition of x to the two modes of representation of t. If n is even, say $2m$, it is clear that

$$b_1 = b_1', \; b_2 = b_2', \ldots b_m = b_m',$$

also
$$b_{m+1} = k^{a_2 + a_4 + \ldots + a_{2m}}(2), \qquad b'_{m+1} = k^{a_2 + a_4 + \ldots + a_{2m} + 1}(0),$$
$$b_{m+2} = k^{a_2 + a_4 + \ldots + a_{2m} + 2}(2), \qquad b'_{m+2} = k^{a_2 + a_4 + \ldots + a_{2m} + 1}(0),$$
$$\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots$$

hence
$$b_{m+1} = b'_{m+1}, \quad b_{m+2} = b'_{m+2}, \ldots;$$

and thus x has the same value whichever of the two forms for t is employed; the case in which n is odd may be similarly treated.

The same result can readily be shewn to hold for y. Therefore, corresponding to any assigned t, x and y are uniquely determined.

Next, let us suppose x and y to be assigned. We have

$$a_1 = b_1, \; a_2 = k^{b_1}(c_1), \; a_3 = k^{c_1}(b_2), \; a_4 = k^{b_1 + b_2}(c_2), \ldots$$
$$a_{2n-1} = k^{c_1 + c_2 + \ldots + c_{n-1}}(b_n), \qquad a_{2n} = k^{b_1 + b_2 + \ldots + b_n}(c_n);$$

for, if $p = k^r(q)$, then $p + q$ is an even number.

In case x, y are both of the second class, t is uniquely determined.

If x is of the first class, and y of the second; let

$$x = \cdot b_1 b_2 \dots b_n 222 \dots = \cdot b_1 b_2 \dots \overline{b_n + 1}000 \dots,$$

$$y = \cdot c_1 c_2 \dots c_n c_{n-1} \dots,$$

and let the two values of t be denoted by $\cdot a_1 a_2 a_3 \dots$, $\cdot a_1' a_2' \dots$.

It is clear that

$$a_1 = a_1',\ a_2 = a_2',\ \dots\ a_{2n-1} = a'_{2n-1};$$

also

$$a_{2n} = k(a'_{2n}),\ a_{2n+1} = k^{c_1+c_2+\dots+c_n}(b_n),\ a'_{2n+1} = k^{c_1+c_2+\dots+c_n}(b_n + 1);$$

thus a_{2n+1}, a'_{2n+1} are not identical, although a_{2n}, a_{2n}' will be so if each is unity. It is thus seen that t has two distinct values corresponding to one point (x, y), when x is a number of the first class, and y is of the second class. It can be shewn in a similar manner that there are four points t corresponding to a single point (x, y) such that x, y are both numbers of the first class.

The correspondence is continuous. For if t, t' are identical as regards the first $2n$ figures, x and x' are identical as regards their first n figures, and the same is true of y and y'.

The curve which has thus been defined is a continuous curve which passes through each point in the square at least once; there is an everywhere dense enumerable set of points through each of which the curve passes twice, and another everywhere dense enumerable set of points through each of which it passes four times; through each point of the remaining unenumerable set of points, the curve passes once only.

The plane measure of an arc of Peano's curve which corresponds to an interval (t_0, t_1) is not zero, *i.e.* the area which a number of rectangles enclosing all the points of the arc have in common has a lower limit greater than zero.

The two continuous functions $f(t)$, $\phi(t)$, which define x, y as functions of t, do not possess, for any value of t, definite differential coefficients, and are perhaps the simplest examples of continuous non-differentiable functions.

327. It might at first sight appear that a curve having the same properties as that of Peano might have been defined by restricting $t = \cdot a_1 a_2 \dots$ to be such that an infinite number of digits other than 0 are present, and then defining x, y by

$$x = \cdot a_1 a_3 a_5 \dots,\ y = \cdot a_2 a_4 a_6 \dots.$$

If however the double representation of x, y were not restricted, as in the case of t, there would be no value of t corresponding to, say,

$$x = \cdot 1000 \dots,\ y = \cdot 2000 \dots.$$

If (x, y) were on the other hand so restricted, there would be no values of (x, y) corresponding, for example, to

$$t = \cdot 1\,1\,1\,0\,1\,0\,1\,0\,1\,\ldots.$$

It thus appears that some such rule as that given by Peano is necessary to obviate the difficulty caused by the double representation of a certain class of rational numbers, in a given scale.

The method may easily be extended to obtain a continuous correspondence between the points in a cube and those in a linear interval.

A somewhat different method of establishing correspondence between the points of the square, and those of the linear interval, is the following*.

Let t_1 denote one of the perfect set of points defined by

$$t_1 = \frac{a_1}{3} + \frac{a_2}{3^2} + \frac{a_3}{3^3} + \ldots,$$

when every a is either 0 or 2. For such a point t_1, x and y may be defined by

$$x = \frac{1}{2}\left(\frac{a_1}{2} + \frac{a_3}{2^2} + \frac{a_5}{2^3} + \ldots\right),$$

$$y = \frac{1}{2}\left(\frac{a_2}{2} + \frac{a_4}{2^2} + \frac{a_6}{2^3} + \ldots\right).$$

A point t which does not belong to the perfect set is interior to one of the complementary intervals (t_1', t_1'') of the set: in such an interval we may define x, y as linear functions of t, thus

$$x = x' + \frac{x' - x''}{t_1' - t_1''}(t - t_1'),$$

$$y = y' + \frac{y' - y''}{t_1' - t_1''}(t - t_1'),$$

where (x', y'), (x'', y'') correspond to t_1', t_1'' respectively.

328. A method of constructing a continuous curve which fills a square has been given in a geometrical form by Hilbert†.

Divide the interval (0, 1) into four equal parts, and number them in order as 1, 2, 3, 4. Then divide the square into four equal parts, as in Fig. 7, and number them 1, 2, 3, 4, to correspond with the segments of the linear interval. Next divide each segment of the straight line into four equal parts, and each of the four squares into four equal parts as in Fig. 8. The sixteen squares so formed are then numbered in order, so that each square has one side in common with the one next in order; the squares then correspond with the segments numbered in the same way. At the next stage there are (Fig. 9) 64 squares corresponding to 64 segments of the interval (0, 1). Proceeding in this manner indefinitely, any point of

---

* See Lebesgue, *Leçons sur l'intégration* (1904), p. 44.

† See *Math. Annalen*, vol. XXXVIII (1891), p. 459.

0, 1) is determined by the intervals of the successive sets of sub-divisions in which it lies. The corresponding point in the square area is determined by the succession of squares, each containing the next, in which it lies.

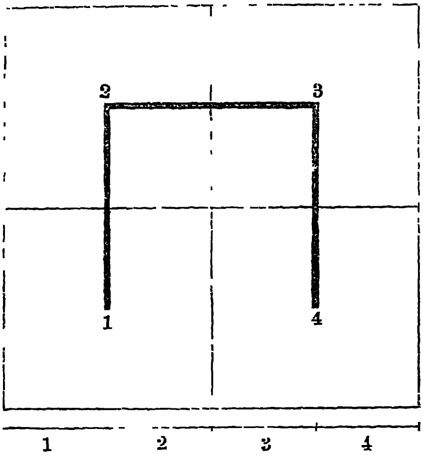

FIG. 7

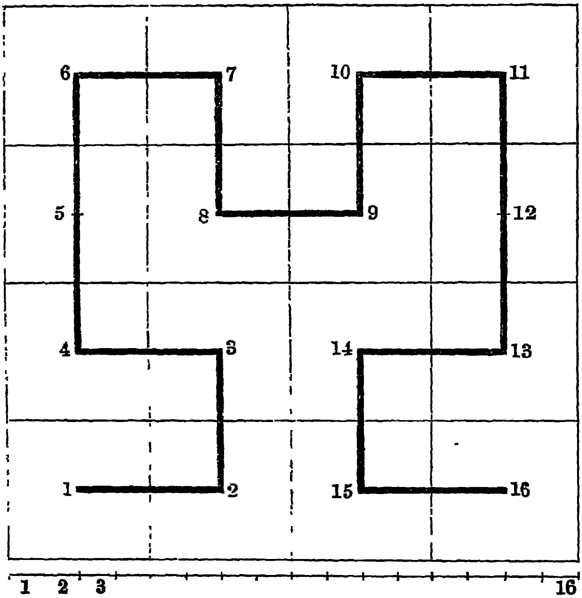

FIG. 8

The curve is thus determined as the limit of a sequence of polygons denoted by the thickened lines in the figures. The curve thus obtained is continuous, but has no tangent. Hilbert remarks that, if the interval

(0, 1) be taken as a time interval, a kinematical interpretation of the functional relation between the curve and the segment is that a point may move so that in a finite time it passes through every point of the square area.

Continuous curves of this kind can be constructed by any method by which an everywhere dense enumerable set of points in the square can be made to correspond with a similar set of points in the linear interval; provided the functional relation $x = f(t)$, $y = \phi(t)$, in such correspondence, is uniformly continuous. For, when this condition is satisfied, the functions obtained by the method of extension of $f(t)$, $\phi(t)$ to the remaining points of (0, 1) as secondary points (see § 287) will yield a correspondence of all the points of the square with those of the linear interval, of the required character.

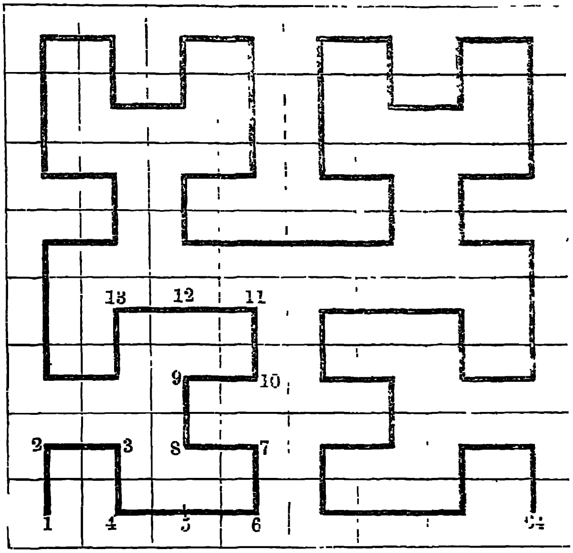

FIG. 9

Another method, differing from that of Hilbert, has been given by E. H. Moore* and by Schoenflies†.

Let m be an uneven number (in the figure, $m = 3$): divide the linear interval (0, 1) into m^2 equal parts, and also the square into m^2 equal parts. Let these squares be passed through by a polygonal line, of which the sides are diagonals of the squares, as in the figure: in this manner the squares are arranged in order 1, 2, 3, ... m^2, and are placed into correspondence with the segments bearing the same numbers. At the same

---

* *Trans. Amer. Math. Soc.* vol. I (1900), p. 77.

† *Bericht über die Mengenlehre*, vol. I, p. 121.

time the end-points of a diagonal so traversed are made to correspond with the end-points of a segment of the linear interval. Thus $m^2 + 1$ points in the linear interval are placed into correspondence with points in the square, so that to each of the $m^2 + 1$ points of the linear interval there is one point in the square; but the converse is not the case. Next, divide each of the m^2 linear intervals into m^2 equal parts, and the corresponding squares into m^2 equal parts: then construct as before a polygon traversing diagonals of all the m^4 squares, making their end-points correspond to the end-points of the corresponding m^4 parts of the linear interval.

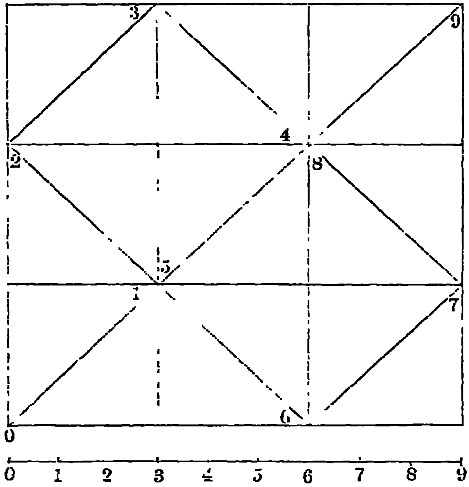

FIG. 10

Proceeding in this manner, we gradually place points in the square, consisting of an everywhere dense enumerable set, into correspondence with a set in the linear interval which possesses the same property; and the functional relation so set up is uniformly continuous. The definition of the functions for the whole linear interval is then obtained, as explained above, by the method of extension. The case $m = 3$ corresponds to Peano's analytical method. In the method of Moore and Schoenflies, the curve is determined as the limit of a sequence of polygons inscribed in the curve. In Hilbert's method the polygons which approximate to the form of the curve are not inscribed in the curve, but are otherwise determined.

CHAPTER VI

THE RIEMANN INTEGRAL

329. The fundamental operation of the calculus, known as integration, regarded from one point of view, consists essentially in the determination of the limit of the sum of a finite series of numbers, as the number of terms of the series is indefinitely increased, whilst the numerically greatest of the individual terms of the series approaches the limit zero. The laws which regulate the specification of the terms of the series must be supposed, in any given instance, to be assigned, and to be of such a character that the limit in question exists. It is in this form that the problem of integration naturally presents itself in ordinary problems of a geometrical or physical character, such as the determination of lengths, areas, volumes, &c. The method of integration, so regarded, has its origin in the method of exhaustions employed by the Greek geometers, and was developed later in forms whose theoretical exactitude depended at various epochs upon the stage which the development of Analysis in general had reached. In the hands of Cauchy, Dirichlet, and Riemann, the definition of the definite integral gradually attained to an exact arithmetic form, which fitted it for the purposes of modern analysis; and in fact the definition given by Riemann leaves nothing to be desired as regards precision. Riemann not only formulated a rigorous arithmetical definition of the integral of a bounded function, but also established a necessary and sufficient condition for the existence of the definite integral of the function. This definition of the definite integral was, in the latter part of the nineteenth century, the one that was employed in all rigorous mathematical analysis, but in the present century it has, for the purposes of theoretical investigations, been largely superseded by the more general formulation of Lebesgue. In accordance with Lebesgue's definition, integrable functions, *i.e.* functions which possess a definite integral, form a class which is markedly wider than, and includes, the class of functions that are integrable in accordance with Riemann's definition. Although the definition, of which the most precise formulation is that of Riemann, is primarily applicable only to a bounded function over a bounded interval, it was extended by Cauchy, Harnack, de la Vallée Poussin, and others, to cases in which the function to be integrated is unbounded, and to integration over unbounded intervals. It was also extended to the case of double, or multiple, integration.

The process of integration, leading to the indefinite integral, is also regarded as the operation inverse to that of differentiation; and the relation of this mode of regarding integration with the one referred to above has been formulated in what is known as the fundamental theorem of the Integral Calculus. Many important investigations are concerned with the relation between the two modes of regarding integration, and with the establishment of the fundamental theorem, including an examination of the limitations to which it is subject. It is in this connection that the advantages of the Lebesgue Integral and its generalizations, over the Riemann Integral, are most apparent.

The Riemann Integral is not only of interest from an historical point of view, but it still possesses great intrinsic importance in Analysis, and will probably continue to be the basis upon which practical applications of the Integral Calculus rest. Accordingly an account of the theory of Riemann integration will be given in the present Chapter. An account is also given in the present Chapter of the properties of the Riemann Integral, and of some of its extensions to the case of unbounded functions, or of integration over unbounded domains. A more complete account of these extensions is however included in the more general theory of Lebesgue integration, and its extensions, which will be discussed in Chapters VII and VIII.

THE RIEMANN INTEGRAL IN A LINEAR INTERVAL

330. Let $f(x)$ be a bounded function, defined for the closed linear interval (a, b), where $b > a$: so that there exist an upper boundary U, and a lower boundary L, of the functional values in the closed interval. Let a system of nets, with closed meshes, be applied to the interval (a, b), and let $\delta_1^{(n)}, \delta_2^{(n)}, \dots \delta_{m_n}^{(n)}$ denote the breadths of the m_n meshes of the net D_n, of the system. Let $M(\delta_r^{(n)})$ denote any number so chosen as to be not greater than the upper boundary of the function $f(x)$ in the closed mesh $\delta_r^{(n)}$, and so as not to be less than the lower boundary of $f(x)$ in the same mesh, and consider the sum

$$S_n = \delta_1^{(n)} M(\delta_1^{(n)}) + \delta_2^{(n)} M(\delta_2^{(n)}) + \dots + \delta_{m_n}^{(n)} M(\delta_{m_n}^{(n)}).$$

If the sequence $S_1, S_2, \dots S_n, \dots$ be convergent, and have the same number S for its limit, whatever system of nets, applied to (a, b), be employed, and however the numbers $M(\delta_r^{(n)})$ be chosen, subject only to their limitation in relation to the upper and lower boundaries of $f(x)$ in the meshes $\delta_r^{(n)}$, then the function $f(x)$ is said to have a Riemann integral in the interval (a, b), and the number S defines the value of its integral. The integral, when the limit S exists, is denoted by $\int_a^b f(x)\,dx$.

It will be observed that $M(\delta)$ is not necessarily the value of $f(x)$ at any point in the interval δ; for all that is necessary is that it should not be greater than the upper boundary, nor less than the lower boundary, of $f(x)$ in the closed interval δ. In this respect the definition has a slightly different form from that given by Riemann*, who took $M(\delta)$ to be the value of $f(x)$ at some point in the interval.

It can easily be seen that this alteration in the form of the definition is unessential. Let $U(\delta_r^{(n)})$, $L(\delta_r^{(n)})$ denote the upper and the lower boundary of $f(x)$ in the closed mesh $\delta_r^{(n)}$, and let ϵ_n denote a [illegible] sequence of positive numbers converging to zero. A point $\xi_r^{(n)}$ in the mesh $\delta_r^{(n)}$ exists such that $U(\delta_r^{(n)}) - f(\xi_r^{(n)}) < \epsilon_n$; and we then have

$$\sum_{r=1}^{r=m_n} \delta_r^{(n)} U(\delta_r^{(n)}) - \sum_{r=1}^{r=m_n} \delta_r^{(n)} f(\xi_r^{(n)}) < \epsilon_n (b-a).$$

From this it follows that if either of the limits

$$\lim_{n\sim\infty} \sum_{r=1}^{r=m_n} \delta_r^{(n)} U(\delta_r^{(n)}),\quad \lim_{n\sim\infty} \sum_{r=1}^{r=m_n} \delta_r^{(n)} f(\xi_r^{(n)})$$

exists, then the other exists, and the two have the same value. Similarly it can be seen that a corresponding result holds when $L(\delta_r^{(n)})$ is substituted for $U(\delta_r^{(n)})$, the points $\xi_r^{(n)}$ being then so chosen that

$$f(\xi_r^{(n)}) - L(\delta_r^{(n)}) < \epsilon_n.$$

If then $\lim\limits_{n\sim\infty} \sum\limits_{r=1}^{r=m_n} \delta_r^{(n)} f(\xi_r^{(n)})$ exists, and is independent of the mode in which the points $\xi_r^{(n)}$ of the intervals $\delta_r^{(n)}$ are chosen, and of the system of nets, it follows that the limits of $\sum\limits_{r=1}^{r=m_n} \delta_r^{(n)} U(\delta_r^{(n)})$ and $\sum\limits_{r=1}^{r=m_n} \delta_r^{(n)} L(\delta_r^{(n)})$ exist, and consequently that of $\sum\limits_{r=1}^{r=m_n} \delta_r^{(n)} M(\delta_r^{(n)})$, exist and have the same value independent of the particular system of nets.

A Riemann integral will be spoken of as an *R*-integral, and a function $f(x)$ which has an *R*-integral, in (a, b), will be said to be integrable (R) in (a, b).

The definition of a definite integral, of which Riemann's definition is a development, was given by Cauchy for the case of a continuous function. Cauchy's definition† is in fact what arises when $M(\delta)$ is in every case restricted to be the functional value at one end of the interval δ. Thus it may be expressed by

$$\int_a^b f(x)\,dx = \lim\,[(x_1 - a) f(a) + (x_2 - x_1) f(x_1) + \ldots + (b - x_{n-1}) f(x_{n-1})],$$

where $a, x_1, x_2, \ldots x_{n-1}, b$ are the end-points of a set of sub-divisions of (a, b), and the limit is determined under the conditions stated above, which involve the convergence to zero of the length of the greatest of the sub-intervals $(a, x_1), (x_1, x_2), \ldots (x_{n-1}, b)$.

---

* See Riemann's *Gesam. Werke*, 2nd ed. p. 239.

† *Journal de l'école polytechnique*, cah. XIX (1823), pp. 571 and 590.

THE UPPER AND LOWER RIEMANN INTEGRALS

331. The investigation of the necessary and sufficient conditions that the bounded function $f(x)$ may have an R-integral in (a, b), in accordance with the above definition, is considerably simplified by the introduction of the upper and lower R-integrals of the function $f(x)$ in the interval (a, b).

If (α, β) be any interval contained in (a, b), the upper and lower boundaries of $f(x)$ in the closed interval (α, β) may be denoted by U_α^β, L_α^β respectively. The upper and lower boundaries of $f(x)$ in (a, b) are accordingly denoted by U_a^b, L_a^b; or simply by U and L. If the interval (α, β) be denoted by a single letter δ, we may write $U(\delta)$, $L(\delta)$ for U_α^β, L_α^β respectively.

If (a, b) *be divided into any number of parts denoted by* $\delta_1, \delta_2, \dots \delta_m$, *taken in order from left to right, the sum* $\sum\limits_{r=1}^{r=m} \delta_r U(\delta_r)$ *has a definite lower boundary, and the sum* $\sum\limits_{r=1}^{r=m} \delta_r L(\delta_r)$ *has a definite upper boundary, when all possible modes of dividing* (a, b) *into parts are taken into account. These lower and upper boundaries are defined to be the upper, and the lower*, R-integral of* $f(x)$ *in* (a, b), *and are denoted by* $\overline{\int_a^b} f(x)\,dx$, $\underline{\int_a^b} f(x)\,dx$ *respectively.*

In each of the sums, δ_r denotes the length of the interval described by the same letter. Since

$$\sum_{r=1}^{r=m} \delta_r U(\delta_r) \geq L(b-a), \quad \sum_{r=1}^{r=m} \delta_r L(\delta_r) \leq U(b-a),$$

it is clear that the first sum has a finite lower boundary, and that the second sum has a finite upper boundary. The upper and lower integrals consequently always exist.

The following theorem will be established:

If ϵ *be an arbitrarily chosen positive number, a number d can be so determined that, for any net with closed meshes such that the breadths of all its meshes are* $< d$, *the sum* $\sum\limits_{r=1}^{r=m} \delta_r U(\delta_r)$ *exceeds the upper integral of the function by less than* ϵ.

---

* The upper integral and the lower integral are named by Jordan "l'intégrale par excès" and "l'intégrale par défaut" respectively; see his *Cours d'Analyse*, vol. I, p. 34. They were introduced by Darboux, *Annales de l'école normale*, (2), vol. IV (1875), p. 72, and also by Thomae, *Einleitung* (1875), p. 12; and by Ascoli, *Atti dei Lincei*, (2), vol. II (1875), p. 863.

Since $\overline{\int_a^b} f(x)\,dx$ is the lower boundary of the sum $\sum_{r=1}^{r=m} \delta_r U(\delta_r)$ for all possible nets, a net D exists, such that $\sum_{r=1}^{r=m} \delta_r U(\delta_r) < \overline{\int_a^b} f(x)\,dx + \frac{1}{2}\epsilon$; where the summation is taken for the m meshes of the net D. Let d be chosen to be such that $2mUd < \frac{1}{2}\epsilon$; and consider any net for which all the meshes have breadths $< d$. Those meshes of this second net which contain, in their interiors or at their ends, end-points of the meshes of D are at most $2m$ in number, and the part of the sum $\Sigma\,\delta U(\delta)$, taken for the second net, which corresponds to these $2m$ meshes is $< \frac{1}{2}\epsilon$. All the other meshes of the second net are interior to meshes of D, and the part of $\Sigma\,\delta U(\delta)$ which corresponds to these meshes is less than $\overline{\int_a^b} f(x)\,dx + \frac{1}{2}\epsilon$. Therefore the sum $\Sigma\,\delta U(\delta)$, for all the meshes of the second net, is $< \overline{\int_a^b} f(x)\,dx + \epsilon$; and this is the case for every net the breadths of whose meshes are all $< d$. Also we have $\Sigma\,\delta U(\delta) \geq \overline{\int_a^b} f(x)\,dx$.

For any system of nets $\{D_n\}$, the breadths of all the meshes of D_n are $< d$, for all values of n, from and after some fixed integer. We thus have the following theorem:

If a system of nets, with closed meshes, be fitted on to (a, b), and Σ_n denote the sum

$$\delta_1^{(n)} U(\delta_1^{(n)}) + \delta_2^{(n)} U(\delta_2^{(n)}) + \dots + \delta_{m_n}^{(n)} U(\delta_{m_n}^{(n)}),$$

where $\delta_1^{(n)}, \delta_2^{(n)}, \dots \delta_{m_n}^{(n)}$ denote the breadths of the m_n meshes of the nth net D_n, of the system; then $\Sigma_1, \Sigma_2, \dots \Sigma_n, \dots$ is a sequence of numbers which converges to $\overline{\int_a^b} f(x)\,dx$, whatever be the system of nets employed.

The corresponding theorem for $\underline{\int_a^b} f(x)\,dx$ may be proved in a similar manner; or it may be deduced from the above by observing that

$$\overline{\int_a^b} \{-f(x)\}\,dx = -\underline{\int_a^b} f(x)\,dx.$$

It can be at once deduced from these theorems that

$$\overline{\int_a^b} f(x)\,dx \geq \underline{\int_a^b} f(x)\,dx.$$

For we need only consider a single system of nets. Since, in any net of the system, $\Sigma_n \geq \Sigma_n'$, where Σ_n' denotes the sum

$$\delta_1^{(n)} L(\delta_1^{(n)}) + \delta_2^{(n)} L(\delta_2^{(n)}) + \dots;$$

it follows, by letting n be indefinitely increased, that the upper integral is not less than the lower integral.

If, in the sum Σ_n, we had taken, instead of $U(\delta)$, the upper boundary of $f(x)$ in the open interval δ, we should have obtained the sum

$$\underline{\Sigma}_n = \sum_{r=1}^{r=m_n} \delta_r^{(n)} \, U_{x_{r-1}+0}^{x_r-0},$$

where the interval $\delta_r^{(n)}$ is (x_{r-1}, x_r), and $U_{x_{r-1}+0}^{x_r-0}$ denotes the upper boundary of $f(x)$ in the open interval $\delta_r^{(n)}$.

We have $\Sigma_n \geq \underline{\Sigma}_n$; hence Σ and $\underline{\Sigma}$, the lower boundaries of Σ_n, $\underline{\Sigma}_n$, satisfy the condition $\Sigma \geq \underline{\Sigma}$.

Let Σ_{n+p} be compared with $\bar{\Sigma}_n$. In any mesh δ', of D_{n+p}, that is contained in (x_{r-1}, x_r), or δ, the value of $U(\delta')$ is $\leq U_{x_{r-1}+0}^{x_r-0}$; unless δ' has an end-point at x_{r-1} or x_r, in which case the upper boundaries of $f(x)$ in the closed and in the open interval δ differ from one another by not more than $U - L$. Hence we have $\Sigma_{n+p} - \underline{\Sigma}_n \leq 2(U - L)\,\bar{\delta} m_n$, where δ is the length of the greatest mesh in D_{n+p}. As p is increased indefinitely, $\bar{\delta}$ is diminished indefinitely, and therefore $\Sigma \leq \underline{\Sigma}_n$. As this holds for each value of n, we have $\Sigma \leq \underline{\Sigma}$; and since also $\Sigma \geq \underline{\Sigma}$, we have shewn that Σ and $\underline{\Sigma}$ are equal. It is clear that the corresponding result will hold in the case of the lower integral. It has thus been shewn that:

In the definitions of the upper and lower integrals of a function the upper and lower boundaries of the function in an open interval δ may be employed instead of the upper and lower boundaries in the corresponding closed interval.

332. It has thus been shewn that a bounded function $f(x)$, defined for the interval (a, b), always possesses an upper, and a lower, integral in that interval. The necessary and sufficient condition that $f(x)$ should have an R-integral in (a, b) is that the upper, and the lower, integrals in the interval should be equal. It has been shewn that the upper and lower integrals are the limits of $\sum_{r=1}^{r=m_n} \delta_r^{(n)} U(\delta_r^{(n)})$, $\sum_{r=1}^{r=m_n} \delta_r^{(n)} L(\delta_r^{(n)})$, for a particular system of nets, and are independent of that system.

That the condition is necessary follows from the fact that all the numbers $M(\delta)$, in the sums $S_1, S_2, \ldots$, may be made identical with $U(\delta)$; or that all may be made identical with $L(\delta)$. That the condition is sufficient follows from the fact that S_n lies between

$$\sum_{r=1}^{r=m_n} \delta_r^{(n)} U(\delta_r^{(n)}) \text{ and } \sum_{r=1}^{r=m_n} \delta_r^{(n)} L(\delta_r^{(n)});$$

and thus that, when the two latter sums have the same limit, that limit is also the limit of S_n. We have therefore obtained the following theorem:

The necessary and sufficient condition that the bounded function $f(x)$ may be integrable (R), in (a, b), is the following: Let $F(\delta^{(n)})$ denote the*

---

* See Riemann's *Gesam. Werke*, 2nd ed. p. 240.

fluctuation $U(\delta^{(n)}) - L(\delta^{(n)})$ *of the function in the interval* $\delta^{(n)}$, *which may be either closed or open; then it must be possible to define a system of nets fitted on to* (a, b), *such that, for the net* D_n, *the sum*

$$\delta_1^{(n)}F(\delta_1^{(n)}) + \delta_2^{(n)}F(\delta_2^{(n)}) + \ldots + \delta_{m_n}^{(n)}F(\delta_{m_n}^{(n)})$$

has the limit zero, as n is increased indefinitely.

That this limit exists, and is equal to zero, is equivalent to the statement that, corresponding to an arbitrarily chosen positive number ϵ, a net D_n belonging to a given system, exists, such that the absolute value of

$$\sum_{r=1}^{r=m_n} \delta_r^{(n)}F(\delta_r^{(n)}),$$

for that value of n, and for all greater values, is less than ϵ.

The necessary and sufficient condition for the existence of $\int_a^b f(x)\,dx$ may be stated in the following somewhat more convenient form:

If any system of nets be fitted on to the interval (a, b), *and* k *be an arbitrarily chosen positive number, the sum of the lengths of those meshes of* D_n *in which the fluctuation of* $f(x)$ *is* $\geq k$ *must converge to zero, as n is indefinitely increased.*

To see that the condition so stated is sufficient, we observe that, if $s^{(n)}$ be the sum of the lengths of those meshes of D_n in which the fluctuation of $f(x)$ is $\geq k$, then $\sum_{r=1}^{r=m_n} \delta_r^{(n)}F(\delta_r^{(n)}) \leq s^{(n)}(U - L) + k(b - a - s^{(n)})$.

Since $s^{(n)} \sim 0$, as $n \sim \infty$, the upper limit of the sum on the left-hand side is $\leq k(b - a)$. Since k is arbitrary, the sum converges to zero.

To shew that the condition is necessary, we observe that

$$\sum_{r=1}^{r=m_n} \delta_r^{(n)}F(\delta_r^{(n)}) \geq ks^{(n)} + (b - a - s^{(n)})\bar{F} \geq ks^{(n)};$$

where $\bar{F}$ is the least of the fluctuations in all the intervals $\delta^{(n)}$. Unless $s^{(n)}$ converges to zero, it is impossible that $\Sigma\delta_r^{(n)}F(\delta_r^{(n)})$ can have the limit zero.

The necessary condition for the existence of the R-integral may, in view of the first theorem of § 331, be stated as follows:

The necessary and sufficient condition that the bounded function $f(x)$ *may be integrable* (R) *in* (a, b), *is that, corresponding to an arbitrarily chosen positive number* ϵ, *a positive number* d *can be so determined that, for every net such that the maximum breadth of its meshes is* $< d$, *the sum* $\sum_{r=1}^{r=m} \delta_r F(\delta_r)$ *is less than* ϵ.

333. The most succinct form of the necessary and sufficient condition that a bounded function is integrable (R) is the following*:

The necessary and sufficient condition that a bounded function may be

---

* Lebesgue, *Annali di Mat.* (3)ª, vol. VII (1902), p. 254.

integrable (*R*), *in the interval for which it is defined, is that the points of discontinuity of the function form a set of measure zero.*

It is convenient to express this condition in the form that the function must be continuous *almost everywhere* in the interval.

To shew that the condition is necessary, let us consider the closed set G_k at which the saltus $\omega(x)$, of $f(x)$, is $\geq k$, where k is a positive number. If an interval δ contain a point of G_k within it, the fluctuation of $f(x)$ in δ is $\geq k$. If a point of G_k is the common end-point of two intervals, of equal length, the fluctuation of $f(x)$ in one at least of these intervals is $\geq \frac{1}{2}k$; hence the part which these two intervals contribute to the sum $\Sigma\delta F(\delta)$ is $\geq \frac{1}{2}k\delta$. If we have a net with equal meshes fitted on to (a, b), the contribution of all those meshes which contain, within them or at an end-point, a point of G_k, is not less than the product of $\frac{1}{4}k$ into the sum of the breadths of these meshes. Unless the content of G_k is zero, the sum of the breadths of these meshes is greater than some fixed positive number, for all the nets of a symmetrical system. It is therefore necessary for the existence of the R-integral that the content of G_k should be zero; and this must be the case for every positive value of k. The set of points of discontinuity of the function is the outer limiting set of $\{G_{k_n}\}$, where $\{k_n\}$ is a sequence of diminishing values of k that converges to zero. It follows that the set of points of discontinuity of the function must have measure zero.

To shew that the condition is sufficient, we observe that, if the content of G_k is zero, all the points of G_k are contained within intervals of a finite set the sum of whose lengths is $< \epsilon$. The intervals complementary to this finite set have a total measure $> b - a - \epsilon$, and at every point in each of them $\omega(x) < k$. In accordance with the theorem of § 234, each of these complementary intervals can be divided into a number of parts, in each of which the fluctuation is $< 2k$. Let this be done for each of the complementary intervals. We have now a net fitted on to (a, b), such that the sum of the breadths of those meshes in which the fluctuation is $\geq 2k$ is $< \epsilon$.

For this net $\Sigma\delta F(\delta) < \epsilon(U - L) + 2k(b - a - \epsilon)$; and since k and ϵ are both arbitrarily small, a net can be determined for which $\Sigma\delta F(\delta)$ has an arbitrarily small value. The condition of integrability is therefore satisfied if, for every value of k, G_k has content zero, that is, if the set of points of discontinuity of the function has measure zero.

334. The following theorem[*] will now be established:

---

[*] This theorem was given by de la Vallée Poussin, see his *Cours d'Analyse*, 3rd ed. vol. I, p. 254. A proof of the theorem, other than that given above, was provided by Pollard, see *Messenger of Math* vol. XLIV (1915), p. 141, where a criticism is given of de la Vallée Poussin's proof.

If $f(x)$ be bounded in (a, b), and $\omega(x)$ be the saltus of $f(x)$, at x, then

$$\overline{\int_a^b} f(x)\,dx - \underline{\int_a^b} f(x)\,dx = \overline{\int_a^b} \omega(x)\,dx.$$

If ξ be any point in (a, b), there exists an interval $(\xi - \eta, \xi + \eta')$ containing ξ as an interior point, such that, in any interval interior to that interval, the fluctuation of $f(x)$ is less than $\omega(\xi) + \epsilon$. If d be an arbitrarily chosen number, and λ be the smallest of the three numbers $\frac{1}{2}d$, $\frac{1}{2}\eta$, $\frac{1}{2}\eta'$, the interval $(\xi - \lambda, \xi + \lambda)$, or the part of it in (a, b), is of length not greater than d, and the fluctuation in it is $< \omega(\xi) - \epsilon$. Taking fixed values of the positive numbers d, ϵ, a single such interval corresponds to each point ξ, of (a, b). Employing the Heine-Borel theorem, it follows that there exists a finite set of overlapping intervals, covering (a, b), such that, if δ be any one of them, the fluctuation of $f(x)$ in δ is $< \omega(\xi) - \epsilon$, where ξ is a definite point interior to δ.

Consider two overlapping intervals (α, β), (α', β'), where $\beta' > \beta$, of this finite set. Let ξ, ξ' be the two definite points interior to these intervals. If neither ξ nor ξ' is in the part (α', β), common to the intervals, we can replace the two intervals by the non-overlapping intervals (α, β), (β, β'). In the first of these the fluctuation of $f(\xi)$ is $< \omega(\xi) + \epsilon$, and in the second it is $< \omega(\xi') + \epsilon$. If (α', β) contains one of the two points ξ, ξ', say ξ', we replace the intervals by (α, α') and (α', β'), which have the same properties as in the last case. If (α', β) contains both the points ξ, ξ', let $\omega(\xi) > \omega(\xi')$; we then take (α, ξ), (ξ, β') in place of (α, β) and (α', β'). In (α, ξ) the fluctuation is $< \omega(\xi) + \epsilon$, and in (ξ, β') it is also $< \omega(\xi) + \epsilon$. We can proceed in this manner with any pair of intervals that overlap one another, and we finally obtain a net, fitted on to (a, b), such that, in any mesh δ, the fluctuation of the function is $< \omega(\xi_\delta) + \epsilon$; where ξ_δ is a definite point within, or at an end of, the mesh δ; moreover the breadths of the meshes do not exceed d. If $\varpi(\delta)$ be the upper boundary of $\omega(x)$ for all interior points of the mesh δ, there is such a point at which $\omega(x) > \varpi(\delta) - \rho$; where ρ is arbitrarily chosen. Therefore the fluctuation in the mesh is $> \varpi(\delta) - \rho$, and hence is $\geqq \varpi(\delta)$, since ρ is arbitrary. It follows that, for the net, $\Sigma\delta F(\delta) \geqq \Sigma\delta\varpi(\delta)$. Taking a sequence of values of d that converges to zero, $\Sigma\delta\varpi(\delta)$ converges to $\overline{\int_a^b} \omega(x)\,dx$; hence

$$\overline{\int_a^b} f(x)\,dx - \underline{\int_a^b} f(x)\,dx \geqq \overline{\int_a^b} \omega(x)\,dx.$$

Again, the fluctuation in δ is $< \omega(\xi) + \epsilon$; and hence, for the net,

$$\Sigma\delta F(\delta) < \Sigma\delta\omega(\xi) + \epsilon(b - a).$$

It follows that $\lim\limits_{d\sim 0} \Sigma\delta F(\delta)$ cannot exceed $\overline{\int_a^b} \omega(x)\,dx + \epsilon(b-a)$; and since ϵ is arbitrary, we have

$$\overline{\int_a^b} f(x)\,dx - \underline{\int_a^b} f(x)\,dx \leqq \overline{\int_a^b} \omega(x)\,dx.$$

It has thus been proved that

$$\overline{\int_a^b} f(x)\,dx - \underline{\int_a^b} f(x)\,dx = \overline{\int_a^b} \omega(x)\,dx.$$

The theorem of § 333 can be deduced from this result. For the condition $\overline{\int_a^b} \omega(x)\,dx = 0$, is equivalent to the condition $\int_a^b \omega(x)\,dx = 0$; and unless the set of discontinuities of the function has measure zero, for some value of k (> 0) the set G_k must have positive content, in which case it is impossible that the integral can have the value zero.

PARTICULAR CASES OF FUNCTIONS THAT ARE INTEGRABLE (*R*)

335. The following classes of bounded functions satisfy the condition of integrability (*R*) which has been expressed in various forms above.

(1). All functions which are continuous in the intervals for which they are defined.

(2). All functions with only a finite number of discontinuities, or with any enumerable set of discontinuities.

(3). Monotone functions, and all functions with bounded variation.

For, as has been shewn in § 243, the points of discontinuity of a function with bounded variation form an enumerable set.

(4). Generally, every point-wise discontinuous function which is such that the closed set of points for which the saltus is $\geqq k$ has content zero, whatever positive value k may have.

Dini* has given the theorem that *a function is integrable* (*R*), *if at all points where the discontinuity is of the second kind, it is so for all such points only on one and the same side of the point; and at these points the function may be continuous on the other side, or may have ordinary discontinuities on that side*. In particular, *any function which has only ordinary discontinuities is integrable* (*R*).

To prove this we observe that it has been proved in § 239 that, for such a function, the set of points for which the saltus is $\geqq k$ has content zero, whatever positive value k may have. Therefore the condition of integrability is satisfied.

---

* See *Grundlagen*, p. 335.

Riemann's definition of an integral, and the condition for the existence of the integral, are applicable, without essential change, to the case of a function which, for particular values of the variable, has indeterminate functional values lying, in the case of each such point, between finite limits of indeterminancy. At each point of indeterminancy of the function, it is immaterial whether the function be capable of having all, or only some, values between the limits of indeterminancy; thus there is no loss of generality, if the function be regarded as having two values only at each such point, viz. the two limits of indeterminancy at the point. In estimating the fluctuation of the function in a prescribed interval, the upper boundary is found by taking the upper limits of indeterminancy of the function at the special points as functional values at those points, whilst the lower boundary is found by taking the lower limits of indeterminancy at the special points as the functional values at those points. As in the case of a function which is everywhere single-valued, the saltus at any point is defined as the limit of the fluctuation in a neighbourhood of the point, when that neighbourhood is diminished indefinitely. The conditions of integrability are exactly the same as for a function which is everywhere single-valued, viz. that the function be bounded in its domain, and that the set of points of discontinuity of the function have measure zero; or, in other words, that it be bounded, and continuous almost everywhere, in the domain.

EXAMPLES

1. Riemann's function $f(x) = \frac{(x)}{1^2} + \frac{(2x)}{2^2} + \dots + \frac{(nx)}{n^2} + \dots$, where (x) denotes the positive or negative excess of x over the nearest integer, and $(x) = 0$ when x is half-way between two integers, has been shewn in Example 2, § 240, to be point-wise discontinuous, with all its discontinuities ordinary ones, and everywhere dense in the interval $(0, 1)$. Since all the discontinuities are ordinary ones, and the function is bounded, $f(x)$ is integrable in $(0, 1)$.

2. Let $f(x)$ be defined for the interval $(0, 1)$ as follows:—If x be irrational, let $f(x) = 0$; if $x = p/q$, where p/q is in its lowest terms, let $f(x) = 1/q$; also let $f(0) = f(1) = 0$. This function is an integrable point-wise discontinuous null-function; thus $\int_0^1 f(x)\,dx = 0$. There is only a finite number of points at which the functional value exceeds an assigned positive number.

3. Let $f(x) = 0$, for all rational values of x; and $f(x) = 1$, for all irrational values of x. This function is not integrable (R) in any interval, for it is totally discontinuous.

4. Let $f(x)$ be defined[*] for the interval $(0, 1)$ as follows:—For $\frac{1}{2} < x \leqq 1$, let $f(x) = 1$; for $\frac{1}{2^2} < x \leqq \frac{1}{2}$, let $f(x) = \frac{1}{2}$; for $\frac{1}{2^3} < x \leqq \frac{1}{2^2}$, let $f(x) = \frac{1}{2^2}$: and generally, for

$$\frac{1}{2^{n+1}} < x \leqq \frac{1}{2^n},\ \text{let } f(x) = \frac{1}{2^n};\ \text{and } f(0) = 0.$$

---

[*] Dini-Lüroth, *Grundlagen*, p. 344.

This function is integrable, and $\int_0^x f(x)\,dx = \frac{x}{2^{m-1}} - \frac{1}{3.2^{2m-2}}$, where x is between $\frac{1}{2^m}$ and $\frac{1}{2^{m-1}}$.

GEOMETRICAL INTERPRETATION OF RIEMANN INTEGRATION

336. Let $f(x)$ be a bounded function, defined for the interval (a, b), of which the values are all $\geqq 0$. Associated with the function there exists a plane set of points (x, y) consisting of all the points of which the coordinates satisfy the conditions $a \leqq x \leqq b$, $0 \leqq y \leqq f(x)$. In accordance with Jordan's theory of measure of sets of points (see § 142), this set has an exterior extent, and an interior extent; and the set of points is measurable (J) when the two have the same value, in which case their common value is the extent, or measure (J), of the set. The extent of a two-dimensional set of points may be regarded as a generalization of the conception of area; thus in general, the exterior extent and the interior extent may be spoken of as the *exterior area* and the *interior area* of the space bounded by the axis of x, the two straight lines $x = a$, $x = b$, and the "curve" defined by $y = f(x)$. This set of points G has an area, in the ordinary sense, when the exterior area, and the interior area, are equal; in which case the function $y = f(x)$ is said to be *quadrable* in the interval (a, b).

If a fundamental rectangle be taken, which contains the plane set in its interior, we may fit on to this rectangle a system $\{D_n\}$ of nets, with closed meshes. Consider those meshes of D_n which are such that every point of each mesh is an interior point of G. If the segment (a, b), on the x-axis, be divided by means of the boundaries of the meshes of D_n into intervals $\delta_1, \delta_2, \ldots \delta_{\bar{n}}$, we see that the sum of the measures of those meshes of D_n which consist entirely of interior points of G is $\sum_1^{\bar{n}} \delta L(\delta) + \eta_n$, where η_n is a number which converges to zero, as n is indefinitely increased. Similarly, the sum of those meshes of D_n each of which contains at least one point of G, or of the boundary of G is $\sum_1^{\bar{n}} \delta U(\delta) + \rho_n$, where $\rho_n \sim 0$, as $n \sim \infty$.

It thus appears that $\overline{\int_a^b} f(x)\,dx$ is the exterior extent of G, and that $\underline{\int_a^b} f(x)\,dx$ is the interior extent of G.

If $\int_a^b f(x)\,dx$ exists as a definite number, the plane set G is measurable (J), and the value of the integral measures the area bounded by $x = a$, $x = b$, $y = 0$, $y = f(x)$.

It has been shewn, in § 142, that the condition for the measurability (J), of the set G, is that the frontier of G should have the plane measure zero. It is clear that, in any case, the plane measure of the three rectilinear portions of the boundary of G is zero; thus the condition for the existence of the integral is that the set of boundary points which consists of points on the "curve" $y = f(x)$, or of limiting points of points on the curve, shall have plane measure zero. This condition is equivalent to the condition that the linear measure of the set of points of discontinuity of the function $f(x)$ in the linear interval (a, b) is zero.

In case the bounded function $f(x)$ is not everywhere ≥ 0 in (a, b), we may take $f(x) = f_1(x) - f_2(x)$; where $f_1(x) = f(x)$, for all values of x for which $f(x) \geq 0$, and $f_1(x) = 0$, for those values of x for which $f(x)$ is negative. In case the two sets of points (x, y) for which $a \leq x \leq b$, $0 \leq y \leq f_1(x)$, and $a \leq x \leq b$, $0 \leq y \leq f_2(x)$ are both measurable (J), the integral $\int_a^b f(x)\,dx$ is the excess of the measure of the first set over that of the second set: and this may be interpreted as the excess of that part of the area defined by $x = a$, $x = b$, $y = f(x)$ which is above the x-axis, over that part which is below it.

If the two plane sets of points be not measurable (J), the exterior and interior extents of the first set are $\overline{\int_a^b} f_1(x)\,dx$, $\underline{\int_a^b} f_1(x)\,dx$; and those of the second set are $\overline{\int_a^b} f_2(x)\,dx$, $\underline{\int_a^b} f_2(x)\,dx$ respectively. The upper integral $\overline{\int_a^b} f(x)\,dx$ is then the excess of the exterior extent of the set $a \leq x \leq b$, $0 \leqq y \leqq f_1(x)$, over the interior extent of the set $a \leq x \leq b$, $0 \leq y \leq f_2(x)$; whilst the lower integral $\underline{\int_a^b} f(x)\,dx$ is the excess of the interior extent of the first set over the exterior extent of the second set.

The condition that $f(x)$ may be integrable (R) is that the frontier which consists of the set of points $a \leq x \leq b$, $y = f(x)$, when closed by adding the limiting points, has its plane measure zero.

If a linear set of points H be defined on the x-axis, and lie in the interval (a, b), a function $\phi(x)$ may be defined by the rule that $\phi(x) = 1$, if x be a point of H, and $\phi(x) = 0$, if x be a point of $C(H)$. The set H has an exterior linear extent, and an interior linear extent, which are given by $\overline{\int_a^b} \phi(x)\,dx$, $\underline{\int_a^b} \phi(x)\,dx$, respectively, as may be seen by referring to the definitions. For it is easily seen that the exterior and interior linear extents of H are numerically equal respectively to the exterior and interior plane extents of the plane set $a \leqq x \leqq b$, $0 \leqq y \leqq \phi(x)$. When H is measurable (J) the function $\phi(x)$ is integrable (R) in (a, b), and $\int_a^b \phi(x)\,dx$ is the measure (J) of H. This measure may be regarded as a generalization of the notion of the length of a linear interval. The condition that the linear set H is measurable (J) is that its frontier, which consists of those points of H that are limiting points of $C(H)$, and of those points of $C(H)$ that are limiting points of H, should have the measure zero.

PROPERTIES OF THE DEFINITE RIEMANN INTEGRAL

337. We proceed to consider the properties of the integral $\int_a^b f(x)\,dx$, of a bounded function $f(x)$, defined for the interval (a, b), which is such that the condition for the existence of the R-integral is satisfied.

(1). *The integral* $\int_b^a f(x)\,dx$ *is defined as the value of* $-\int_a^b f(x)\,dx$.

(2). *If* $f(x)$ *be integrable* (R) *in* (a, b), *so also is* $|f(x)|$, *and*

$$\left|\int_a^b f(x)\,dx\right| \leqq \int_a^b |f(x)|\,dx.$$

For the fluctuation of $|f(x)|$ in any interval δ cannot exceed that of $f(x)$ in the same interval; hence, if $\delta F(\delta)$, for a system of nets, have the limit zero, when $F(\delta)$ is the fluctuation of $f(x)$ in δ, it has also the limit zero when $F(\delta)$ denotes the fluctuation of $|f(x)|$; and thus the latter function is integrable (R). Again U, the upper boundary of $f(x)$ in δ, cannot numerically exceed U', the upper boundary of $|f(x)|$ in the same interval; thus $|\Sigma U\delta| \leq \Sigma U'\delta$, and hence the absolute value of the limit of $\Sigma U\delta$ is $\leq$ that of $\Sigma U'\delta$.

(3). *If* $f(x)$ *be integrable* (R) *in* (a, b), $f(x-h)$ *is integrable* (R) *in* $(a+h, b+h)$, *and* $\int_a^b f(x)\,dx = \int_{a+h}^{b+h} f(x-h)\,dx$.

This follows at once from the definitions of the integrals as the limits of sums.

(4). *If the values of the integrable function* $f(x)$ *be arbitrarily altered at each point of a measurable set of points* G, *the new function* $\phi(x)$ *so obtained is integrable* (R), *provided it be bounded, and also the measure of the derivative* G' *of the set be zero.*

For the only points of discontinuity of $\phi(x)$ which are not points of discontinuity of $f(x)$ are points of G, or of G', and therefore form a set of measure zero; hence all the discontinuities of $\phi(x)$ form a set of points of measure zero, and $\phi(x)$ is therefore integrable (R), provided it be bounded. In particular, the theorem holds for any reducible set G.

Also, if $\phi(x) = f(x)$, at all points belonging to a set which is everywhere dense in (a, b), then, provided $\phi(x)$ be integrable (R), its integral is identical with that of $f(x)$.

For, in the finite sum $\Sigma\delta M(\delta)$, we may take the value of $M(\delta)$ in any interval δ to be one of the values which the two functions $f(x)$, $\phi(x)$ have in common in that interval; hence the sums may all be chosen so as to be the same for the two functions. Thus, if the functions be both integrable (R), their integrals are identical.

(5). *A function $f(x)$ which is integrable (R), in (a, b), is also integrable (R) in any interval (α, β) contained in (a, b).*

For the measure of the set of points of discontinuity of $f(x)$ in (a, b) being zero, the measure of the set of those points of discontinuity which are in (α, β) is also zero, and thus the function is integrable (R) in (α, β).

If c is any point in (a, b), we have

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

For the two integrals on the right-hand side both exist; also a system of nets can be so chosen that the point c is always an end-point of two of the meshes of each net. If this be done, the sum $\Sigma\delta M(\delta)$, for (a, b), may be divided into two parts, one of which contains all the intervals on the left of the point c, and the other all those on the right of that point; thus

$$\Sigma\delta M(\delta) = \Sigma_1\delta M(\delta) + \Sigma_2\delta M(\delta).$$

The limits of the three sums are the three integrals of $f(x)$ in (a, b), (a, c), and (c, b) respectively; thus the theorem is established.

(6). *If* $f_1, f_2, f_3, \dots f_n$ be a finite number of bounded functions, each of which is integrable (R) in (a, b), and if $F(f_1, f_2, \dots f_n)$ be a continuous function with respect to $(f_1, f_2, \dots f_n)$, then the function F is integrable (R) in (a, b).*

For the only points of discontinuity of the function $F(x)$ are those of the functions $f_1(x)$, $f_2(x)$, $\dots f_n(x)$; hence the set of points of discontinuity of $F(x)$ has measure zero; and thus $F(x)$ is integrable (R), since it is also a bounded function.

Important particular cases of the general theorem are the following:

(a). If $f(x) = f_1(x) + f_2(x) + \dots + f_n(x)$, where all the functions $f_r(x)$ are integrable (R), then $f(x)$ is integrable (R).

It can also be shewn that $\int_a^b f(x)\,dx = \sum_{r=1}^{r=n} \int_a^b f_r(x)\,dx$. For, in any interval δ, the upper boundary of $f(x)$ cannot exceed the sum of the upper boundaries of the functions $f_r(x)$. From this we see that

---

* Du Bois-Reymond, *Math. Annalen*, vol. xx (1882), p. 123. See also W. H. Young, *Quarterly Journal of Math.* vol. xxxv (1904), p. 190.

$$\overline{\int_a^b} f(x)\,dx \leq \sum_1^n \overline{\int_a^b} f_r(x)\,dx,$$

and similarly we have $\underline{\int_a^b} f(x)\,dx \geq \sum_1^n \underline{\int_a^b} f_r(x)\,dx$

(*b*). If $f(x) = f_1(x) . f_2(x) \dots f_n(x)$, where all the functions $f_r(x)$ are integrable (R) in (a, b), then $f(x)$ is also integrable (R) in (a, b).

(*c*). If $f(x)$, $\phi(x)$ be integrable (R) in (a, b), and $|\phi(x)|$ always exceeds some fixed positive number A, so that $\dfrac{f(x)}{\phi(x)}$ is a continuous function of (f, ϕ), then $\dfrac{f(x)}{\phi(x)}$ is integrable (R) in (a, b).

(7). *If two functions $f^+(x)$, $f^-(x)$ be defined as follows:—Let $f^+(x) = f(x)$ for all values of x such that $f(x) > 0$, and let $f^+(x) = 0$, when $f(x) \leq 0$; let $f^-(x) = -f(x)$ for all values of x such that $f(x) < 0$, and $f^-(x) = 0$, when $f(x) \geq 0$; then if $f(x)$ be integrable (R) in (a, b), the functions $f^+(x)$, $f^-(x)$ are integrable (R) in (a, b), and* $\int_a^b f(x)\,dx = \int_a^b f^+(x)\,dx - \int_a^b f^-(x)\,dx$.

For the fluctuation of $f^+(x)$ in any interval δ cannot exceed that of $f(x)$ in the same interval; hence, since $\Sigma\delta F(\delta)$, for $f(x)$, has the limit zero. the corresponding sum for $f^+(x)$ has the limit zero; and thus $f^+(x)$ is integrable. In a similar manner it can be shewn that $f^-(x)$ is integrable.

Since $f(x) = f^+(x) - f^-(x)$, we see from (6) (*a*) that

$$\int_a^b f(x)\,dx = \int_a^b f^+(x)\,dx - \int_a^b f^-(x)\,dx.$$

It should be observed that it is not in general true that, if $f(x)$ be integrable in (a, b), and be expressed as the sum $f_1(x) + f_2(x)$, of two bounded functions, then $f_1(x)$, $f_2(x)$ are also integrable in (a, b). For it is clear that, $f(x)$ being given, we may take for $f_1(x)$ any arbitrarily defined non-integrable function, then $f_2(x)$ is also determinate and non-integrable.

(8). *If $f(x)$, $\phi(x)$ be both integrable (R), and be such that $|f(x)| \leq |\phi(x)|$ for every value of x, then* $\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |\phi(x)|\,dx$.

In particular, *if $\phi(x)$ is constant, and equal to P, the upper boundary of $|f(x)|$ in (a, b), then* $\left|\int_a^b f(x)\,dx\right| \leq P(b-a)$.

For $\int_a^b \{|\phi(x)| - |f(x)|\}\,dx$ is ≥ 0, since, in every interval δ, no value of $|\phi(x)| - |f(x)|$ is negative, and thus the sums of which the integral is the limit are all $\geqq 0$. Also from (2), we have $\left|\int_a^b f(x)\,dx\right| \leqq \int_a^b |f(x)|\,dx$, and this is $\leqq \int_a^b |\phi(x)|\,dx$. The particular case follows by assuming $\phi(x) = P$.

If U, L denote the upper and lower boundaries of $f(x)$ in (a, b), then

$$L(b-a) \leqq \int_a^b f(x)\,dx \leqq U(b-a).$$

For $\Sigma\delta U(\delta)$, $\Sigma\delta L(\delta)$ each lies between $U\Sigma\delta$ and $L\Sigma\delta$, or between $U(b-a)$ and $L(b-a)$; the same must hold of the common limit, which is the integral $\int_a^b f(x)\,dx$.

(9). *If $\eta_1, \eta_2, \dots \eta_n, \dots$ be an enumerable set of non-overlapping intervals contained in (a, b), in descending order of length, then the sum of the integrals of $f(x)$ taken over $\eta_1, \eta_2, \dots \eta_n$, converges to a definite finite limit, as n is increased indefinitely; $f(x)$ being a function which is integrable (R) in (a, b).*

Let us denote by S_n the sum of the integrals of $f(x)$ taken over the intervals $\eta_1, \eta_2, \dots \eta_n$. Since $\eta_1 + \eta_2 + \dots + \eta_n$ increases with n, and is always less than $b - a$, it has a definite limit as n is increased indefinitely. we can therefore choose n so great that $\eta_{n+1} + \eta_{n+2} + \dots + \eta_{n+m} < \epsilon$, for every value of m, where ϵ is an arbitrarily chosen positive number. With this value of n, we see that $|S_{n+m} - S_n| < \epsilon . P$, where P is the upper boundary of $|f(x)|$ in (a, b) If η be an arbitrarily chosen positive number, we can choose ϵ such that $\epsilon < \eta/P$; thus n can be so chosen that

$$|S_{n+m} - S_n| < \eta,$$

and hence S_n has a definite limit, as n is increased indefinitely.

(10). If $f_1(x), f_2(x), \dots f_n(x), \dots$ be a sequence of functions, defined in the interval (a, b), and $f(x)$ be such that, for all values of x in (a, b),

$$|f(x) - f_n(x)| < \epsilon,$$

where ϵ is an arbitrarily chosen positive number, provided n is $\geqq$ a fixed integer n_ϵ, dependent on ϵ; the sequence $\{f_n(x)\}$ is said to *converge uniformly to $f(x)$, in the interval (a, b).*

If a sequence of functions $\{f_n(x)\}$, all integrable (R), in the interval (a, b), converges uniformly, in that interval, to the bounded function $f(x)$, then $f(x)$ is integrable (R), and

$$\int_a^b f(x)\,dx = \lim_{n\sim\infty} \int_a^b f_n(x)\,dx.$$

At almost every point of (a, b), that is, almost everywhere in (a, b) (see § 333), all the functions of the sequence $\{f_n(x)\}$ are continuous; for the set of all those points at which any of them are discontinuous has the measure zero. Let ξ be a point at which all the functions $f_n(x)$ are continuous; if $(\xi - h, \xi + h)$ is a neighbourhood of ξ, the fluctuation of $f_{n_\epsilon}(x)$ in $(\xi - h, \xi + h)$ is $< \epsilon$, if h be sufficiently small; the integer n_ϵ having been so chosen that $|f(x) - f_{n_\epsilon}(x)| < \epsilon$, in (a, b). It follows that the fluctuation of $f(x)$ in $(\xi - h, \xi + h)$ is $< 3\epsilon$. Since ϵ is arbitrary, ξ is a point of continuity of $f(x)$; therefore $f(x)$ is continuous almost everywhere in (a, b), and is therefore integrable (R). Also

$$\left| \int_a^b f(x)\,dx - \int_a^b f_{n_\epsilon}(x)\,dx \right| < \epsilon\,(b - a);$$

and therefore $$\int_a^b f(x)\,dx = \lim_{n\sim\infty} \int_a^b f_n(x)\,dx.$$

R-INTEGRALS OF FUNCTIONS OF TWO OR MORE VARIABLES

338. The Riemann definition of the integral of a bounded function, defined for an interval (a, b), may be extended at once* to the case of a bounded function $f(x^{(1)}, x^{(2)})$ defined in a rectangular cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, or more generally to the case of a bounded function $f(x^{(1)}, x^{(2)}, \ldots x^{(p)})$, of p variables, defined in a p-dimensional cell $(a^{(1)}, a^{(2)}, \ldots a^{(p)}; b^{(1)}, b^{(2)}, \ldots b^{(p)})$.

In the definitions given in §§ 330, 331, we have only to consider systems of nets in two, or in p-dimensions, instead of linear systems.

Thus the upper and lower integrals† of $f(x^{(1)}, x^{(2)})$ in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ may be denoted by

$$\overline{\int}_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

$$\underline{\int}_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

which represent the limits of the sums of the products of the areas of the meshes of a net D_n into the upper, or the lower, boundaries of the function in the corresponding meshes. That these limits exist, and are independent of the particular system of nets, is proved exactly as in § 332.

When the upper and lower integrals have equal values, $f(x^{(1)}, x^{(2)})$ is integrable (R) in the fundamental cell, and their common value is denoted by

$$\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}).$$

---

* This extension was made by H. J. S. Smith, *Proc. Lond. Math. Soc.* (1), vol. VI (1875), p. 152, and also by Thomae, *Einleitung in die Theorie der bestimmten Integrale* (1875), p. 33; also *Schlömilch's Zeitschr.* vol. XXI (1876), p. 224.

† See Jordan's *Cours d'Analyse*, vol. I, p. 34. An elaborate treatment of double integration has been given by Stolz, *Grundzuge*, vol. III, where the triangle or polygon is employed in relation to the measure of sets of points, instead of the rectangle. On this matter see Schoenflies, *Bericht*, vol. I, p. 179.

The necessary and sufficient condition that $f(x^{(1)}, x^{(2)})$ is integrable (R), viz. that the plane measure of its points of discontinuity should be zero, or in other words, that the function should be continuous almost everywhere in the cell, is established in the same manner as in § 333, the proof requiring only a slight modification.

The integral

$$\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}), \text{ where } f(x^{(1)}, x^{(2)}) \geqq 0,$$

in the cell is the measure (J) of the three-dimensional set of points $(x^{(1)}, x^{(2)}, x^{(3)})$ defined by

$$a^{(1)} \leqq x^{(1)} \leqq b^{(1)};\ a^{(2)} \leqq x^{(2)} \leqq b^{(2)};\ 0 \leqq x^{(3)} \leqq f(x^{(1)}, x^{(2)}),$$

it being assumed that this measure (J) exists.

A function $f(x^{(1)}, x^{(2)})$ which takes both signs in the fundamental cell is the difference of two functions $f_1(x^{(1)}, x^{(2)})$, $f_2(x^{(1)}, x^{(2)})$, both of which are $\geqq 0$. If the frontiers of both the three-dimensional sets

$$a^{(1)} \leqq x^{(1)} \leqq b^{(1)};\ a^{(2)} \leqq x^{(2)} \leqq b^{(2)};\ 0 \leqq x^{(3)} \leqq f_1(x^{(1)}, x^{(2)}),$$
$$a^{(1)} \leqq x^{(1)} \leqq b^{(1)};\ a^{(2)} \leqq x^{(2)} \leqq b^{(2)};\ 0 \leqq x^{(3)} \leqq f_2(x^{(1)}, x^{(2)}),$$

have their three-dimensional measures zero, the function $f(x^{(1)}, x^{(2)})$ is integrable (R) in the cell, and the integral is the excess of that of $f_1(x^{(1)}, x^{(2)})$ over that of $f_2(x^{(1)}, x^{(2)})$.

The integral

$$\int_{(a^{(1)},\, a^{(2)},\, \ldots\, a^{(p)})}^{(b^{(1)},\, b^{(2)},\, \ldots\, b^{(p)})} f(x^{(1)}, x^{(2)}, \ldots x^{(p)})\, d(x^{(1)}, x^{(2)}, \ldots x^{(p)})$$

may be considered in a precisely similar manner, and the condition for its existence is an extension of that for the cases $p = 1, 2$.

339. Next, let G be a bounded set of points in plane space; it is therefore contained in the interior of a fundamental rectangular cell $(a^{(1)}, a^{(2)};\ b^{(1)}, b^{(2)})$. Let us suppose that a bounded function $f(x^{(1)}, x^{(2)})$ is defined for all the points of G. Let $\bar{f}(x^{(1)}, x^{(2)})$ be defined in the fundamental cell by the conditions

$$\bar{f}(x^{(1)}, x^{(2)}) = f(x^{(1)}, x^{(2)}), \text{ at all points of } G,$$

and

$$\bar{f}(x^{(1)}, x^{(2)}) = 0, \text{ at all points of } C(G).$$

If we consider, for the function $\bar{f}(x^{(1)}, x^{(2)})$, the sum $\Sigma\delta\,\{U(\delta) - L(\delta)\}$, for those meshes of the plane net D_n, fitted on to the fundamental cell, which contain a point of G and also a point of $C(G)$, we see that, for all such meshes, $U(\delta) - L(\delta) > 0$, unless $U(\delta)$ and $L(\delta)$ are both zero, since there is at least one point in the mesh at which $\bar{f}(x^{(1)}, x^{(2)}) = 0$. Thus $\Sigma\delta\,\{U(\delta) - L(\delta)\}$, taken for all such meshes, is > 0, unless $\bar{f}(x^{(1)}, x^{(2)}) = 0$ at every point of all these meshes.

Unless $f(x^{(1)}, x^{(2)}) = 0$, at every point of the frontier of G, with the possible exception of a part of G of which the measure (J) is zero, the limit of $\Sigma\delta \{U(\delta) - L(\delta)\}$, as $n \sim \infty$, will not be zero unless $\Sigma\delta$, taken for all meshes that contain points on the frontier of G, converges to zero, as $n \sim \infty$. Thus it follows that the upper and lower integrals of $\bar{f}(x^{(1)}, x^{(2)})$ in the fundamental cell will have unequal values, and thus that $f(x^{(1)}, x^{(2)})$ is not integrable (R), unless, either the set G is measurable (J), or

$$f(x^{(1)}, x^{(2)}) = 0$$

at all the points of its frontier with the possible exception of a part, of which the measure (J) is zero.

We define the R-integral

$$\int_{(G)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

of $f(x^{(1)}, x^{(2)})$ over the set G, to be the value of

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \bar{f}(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

when the latter integral exists, which we have seen is only, in general, the case when the set G is measurable (J).

In the definition of the R-integral of a bounded function $f(x^{(1)}, x^{(2)})$, defined for a bounded set G, it will accordingly be assumed that G is measurable (J); and then the integral exists when the function that is equal to $f(x^{(1)}, x^{(2)})$ at all points of G, and is elsewhere zero, has an R-integral in a cell that contains G.

An integral
$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$
or
$$\int_{(G)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

is, in accordance with tradition, spoken of as a double integral, although it is defined as the single limit of a finite sum, and accordingly the sign of integration is here employed only once. The term "double," in the name double integral, must be taken to have reference to the two-dimensional set of points for which the function is defined. A similar remark applies to the case of a p-fold integral. The above notation is such that if

$$(x^{(1)}, x^{(2)}),\ \text{or}\ (x^{(1)}, x^{(2)}, \dots x^{(p)}),$$

be denoted by a single letter x, a double, or p-fold, integral may be denoted by

$$\int_a^b f(x)\, dx,$$

or by

$$\int_{(G)} f(x)\, dx,$$

independently of the number of dimensions of the space in which the points a, b, x lie. This is in accordance with the parity of the properties of the R-integral in any number of dimensions with those of the R-integral in linear space.

340. With slight adaptation, the properties established in § 337 are applicable to the case of functions of two or more variables. The notation will be the same as for functions of one variable, provided the single letter x is employed as typical of $(x^{(1)}, x^{(2)}, \dots x^{(p)})$ in space of p-dimensions.

The definition in (1) of $\int_b^a f(x)\,dx$

as the value of $-\int_a^b f(x)\,dx$

holds for a cell (a, b). The proof of the property (2) is unaltered by employing δ for a cell, instead of a linear interval. The property (3) is unaltered essentially; it becomes, for the case $p = 2$,

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$$
$$= \int_{(a^{(1)}+h^{(1)}, a^{(2)}+h^{(2)})}^{(b^{(1)}+h^{(1)}, b^{(2)}+h^{(2)})} f(x^{(1)} - h^{(1)}, x^{(2)} - h^{(2)})\, d(x^{(1)}, x^{(2)}).$$

The property (4) holds with its form unaltered. The property (5) holds for a cell (α, β) contained in the cell (a, b).

If c is a point in the cell (a, b), the cell (a, b) is divided into 2^p cells by the $(p-1)$-dimensional planes through the point c, and

$$\int_a^b f(x)\,dx$$

is expressed as the sum of the integrals over the 2^p cells into which (a, b) is so divided.

The properties (6), (7), (8) remain unaltered, and (9) holds for an enumerable non-overlapping set of cells in the cell (a, b).

A more general form of (5) which holds for p-ple integrals is the following:

If G be a bounded set that is measurable (J), and if G be the sum of two sets G_1 and G_2, both measurable (J), then

$$\int_{(G)} f(x)\,dx = \int_{(G_1)} f(x)\,dx + \int_{(G_2)} f(x)\,dx.$$

To prove this theorem, let K be the set of points of G at which the saltus of $f(x)$ is $\geqq k$; then K has two components, K_1 a part of G_1, and K_2 a part of G_2. The only points at which the saltus of $f(x)$, considered as defined in G_1 only, and zero elsewhere, is $\geqq k$, consist of those points of K_1 that are interior to G_1, together with points forming a set K_1' on the frontier of G_1 and G_2. Since the content of this frontier is zero, the measure (J) of K_1' is zero; and the measure (J) of the set of points of K_1 interior to G_1 is also zero, since K_1 has the content zero. Since the measure (J) of the set of points at which the saltus of $f(x)$ regarded as defined in G_1, and elsewhere zero, is zero, it follows that $f(x)$ is integrable (R) in G_1; and similarly it can be shewn to be integrable (R) in G_2. The integral

$$\int_{(G)} f(x)\,dx$$

is, by definition, the limit of the finite sum

$$\delta_1 M(\delta_1) + \delta_2 M(\delta_2) + \ldots + \delta_m M(\delta_m)$$

taken for a net fitted on to the fundamental cell, or interval. The meshes δ consist (1), of those which contain interior points of G_1 only, but no points on the frontiers of G_1 and G_2, (2), of those which contain interior points of G_2 only, but no points on the frontiers of G_1 and G_2, and (3), of those which contain points of G_1 and G_2 which are not interior points of either set, but are points on the frontiers of G_1 and G_2. The above sum may be divided therefore into three portions containing those meshes δ which respectively belong to (1), (2), and (3). The limit of the first of these sums is

$$\int_{(G_1)} f(x)\,dx,$$

that of the second is

$$\int_{(G_2)} f(x)\,dx,$$

and that of the third is at most numerically equal to U multiplied by the content of the points on the frontiers of G_1 and G_2; where U denotes the upper boundary of $|f(x)|$ in G. Since the contents of the frontiers are zero, the limit of the third part of the sum is zero. The theorem has accordingly been established.

INTEGRABLE NULL-FUNCTIONS AND EQUIVALENT INTEGRALS

341. If $f(x)$ be integrable (R) in the interval, or cell, (a, b), and be such that in every interval, or cell, contained in (a, b), its integral is zero, then $f(x)$ is said to be an *R-integrable null-function.*

The necessary and sufficient condition that a bounded function $f(x)$ may be an R-integrable null-function is that $f(x) = 0$, almost everywhere in its domain.

Let G_k denote the set which contains the points x at which

$$|f(x)| \geq k,$$

where k is any positive number.

To prove the sufficiency of the condition, let D_n be a net with closed meshes fitted on to the interval, or cell, (a, b), and let $\Sigma\delta'$ denote the sum of the measures of those meshes of D_n which contain at least one point of G_k. If the condition of the theorem is satisfied, $\Sigma\delta'$ has the limit zero, as $n \sim \infty$. The sum $\Sigma\delta M(\delta)$ employed in § 330 is in absolute value

$$< P\Sigma\delta' + (A - \Sigma\delta')\, k,$$

where A is the measure of the interval, or cell, (a, b), and P is the upper boundary of $|f(x)|$ in (a, b). If $\Sigma\delta'$ converges to zero

$$|\Sigma\delta M(\delta)| - Ak$$

is less numerically than the arbitrarily chosen positive number ϵ, for all sufficiently large values of n. If k be chosen $< \epsilon/A$, $|\Sigma\delta M(\delta)| < 2\epsilon$, for large enough values of n. It follows that $\lim\limits_{n\sim\infty} \Sigma\delta M(\delta) = 0$, and therefore $f(x)$ is integrable (R) in (a, b), and its integral has the value zero. The same argument applies to any integral, or cell, (α, β) contained in (a, b).

To shew that the condition is necessary, let it be assumed that $f(x)$ has, in every interval, or cell, contained in (a, b), an R-integral that vanishes. At any point x_1, at which $f(x)$ is continuous, $f(x_1)$ must be zero. For let $f(x_1)$ have, if possible, the positive value c; then a neighbourhood of x_1 can be determined such that, at every point of it, $f(x)$ lies between $c + \epsilon$ and $c - \epsilon$, where ϵ is an assigned positive number $< c$. The integral of $f(x)$ over this neighbourhood, which we take to have the measure λ, is $\geq \lambda(c - \epsilon) > 0$, contrary to the hypothesis. In a similar manner it can be shewn that $f(x_1)$ cannot be negative, therefore $f(x_1) = 0$. Since $f(x)$ is continuous almost everywhere in its domain, it follows that $f(x)$ has the value zero, almost everywhere, that is, with the possible exception of points belonging to a set of measure zero.

342. Two functions $f(x)$, $\phi(x)$ both integrable (R), which have their integrals equal, when taken over any interval, or cell, contained in the fundamental interval, or cell, (a, b), must differ from one another by an R-integrable null-function.

If $f(x)$ be integrable (R), and consequently point-wise discontinuous, and $\psi(x)$ be that function defined, as in § 241, by extension of the function which is defined only at the points of continuity of $f(x)$, and has at those points the same functional values as $f(x)$ itself, then $\psi(x)$ is integrable (R), although it is in general multiple-valued at the points of discontinuity of $f(x)$. It has been explained in § 235 that Riemann's definition is applicable to such a function as $\psi(x)$. That $\psi(x)$ is integrable (R) follows from the fact that its points of discontinuity form a set of which the measure is zero. The function $f(x) - \psi(x)$ is zero at the points of continuity of $f(x)$, and is discontinuous only at the points of discontinuity of $f(x)$, which form a set of measure zero. Consequently $f(x) - \psi(x)$ is an R-integrable null-function, and the two functions have equal integrals in any interval, or cell, for which $f(x)$ is defined. It has thus been shewn that:

A function $f(x)$ that is integrable (R) in an interval, or cell, is the sum of an R-integrable null-function and of the function $\psi(x)$ obtained by extension of the function defined by the values of $f(x)$ at its points of continuity.

THE FUNDAMENTAL THEOREM OF THE INTEGRAL CALCULUS

343. The fundamental theorem of the Integral Calculus asserts that the operations of differentiation and of integration are in general inverse operations. Before we proceed to consider the conditions under which this is the case, the following theorem will be established:

If $f(x)$ be a bounded function which is integrable (R) in the interval (a, b), then $\int_a^x f(x)\,dx$ is a continuous function of x, for the whole interval (a, b), and it is a function of bounded variation in (a, b). It is also absolutely continuous in (a, b).

It has already been shewn that $\int_a^x f(x)\,dx$ exists, for any point x of the interval (a, b); denoting its value by $F(x)$, we have

$$F(x \pm h) - F(x) = \int_x^{x \pm h} f(x)\,dx;$$

hence by (8), of § 337, $| F(x \pm h) - F(x) | \leqq Ph$, where P is the upper boundary of $|f(x)|$ in (a, b). If ϵ be any arbitrarily chosen positive number, and we take $h_1 < \epsilon/P$, then for all values of h which are $\leqq h_1$, we have

$$| F(x \pm h) - F(x) | < \epsilon;$$

but this is the condition of continuity of $F(x)$ at the point x. In case x be one of the end-points of (a, b), h must be restricted to have one sign only.

To prove that $F(x)$ has bounded variation in (a, b), let (a, b) be divided into n sub-intervals by the points $a, x_1, x_2, \ldots x_{n-1}, b$. The sum of the absolute differences of the values of $F(x)$ at the ends of these sub-intervals is

$$\left| \int_a^{x_1} f(x)\,dx \right| + \left| \int_{x_1}^{x_2} f(x)\,dx \right| + \ldots + \left| \int_{x_{n-1}}^{b} f(x)\,dx \right|$$

and this is, in accordance with the theorem (8) of § 337, $\leqq \int_a^b |f(x)|\,dx$; and therefore the sum is less than a fixed positive number. Since the total variation of $F(x)$ in (a, b) is bounded, it follows from the theorem of § 246 that the total fluctuation in the interval is also bounded.

If $(a_1, b_1), (a_2, b_2), \ldots (a_n, b_n), \ldots$ be a finite, or infinite, set of non-overlapping intervals contained in (a, b), of which the total measure is $< \eta$, we have for the finite, or infinite sum

$$\sum_{r=1} | F(b_r) - F(a_r) | \leqq \overline{U} \Sigma (b_r - a_r) < \eta \overline{U},$$

where $\overline{U}$ is the upper boundary of $|f(x)|$ in (a, b). The number η being taken to be $\epsilon/\overline{U}$, where ϵ is an arbitrarily chosen positive number, we see that the condition, given in § 218, for the absolute continuity of $F(x)$ is satisfied.

When $f(x)$ is integrable in (a, b), the function $\int_a^x f(x)\,dx$, which has been shewn to be absolutely continuous, and of bounded variation in (a, b), is said to be the *indefinite integral* corresponding to $f(x)$.

If $f(x)$ be any function defined in (a, b), a function $\phi(x)$ which, at every point x of the interval, possesses a differential coefficient equal to $f(x)$, is said to be a *primitive* of $f(x)$.

The definition is, however, frequently extended to cases in which $\phi'(x)$ either does not exist, or is not equal to $f(x)$, at points belonging to an exceptional set, of measure zero; the condition $\phi'(x) = f(x)$ being satisfied at all points not belonging to the exceptional set.

Taking the function $F(x) = \int_a^x f(x)\,dx$, as the indefinite integral corresponding to $f(x)$, the following properties will be established:

(A). Under certain restrictions $F(x)$ possesses a differential coefficient which is equal to $f(x)$, and thus $F(x)$ is a primitive of $f(x)$.

(B). Also it will be shewn that, if $\phi(x)$ be a function which possesses a differential coefficient $f(x)$, then $f(x)$ may have an indefinite integral $F(x)$, in an interval (a, x), which integral differs from $\phi(x)$ by a constant only; and thus that the indefinite integral of $f(x)$ is determinate except for an additive constant.

It will appear that there are cases of exception to both theorems. When $F(x)$ is an indefinite integral, it may happen that at certain points $F(x)$ does not possess a differential coefficient; and when $\phi(x)$ is a function which possesses a differential coefficient, it is not always the case that the latter is integrable (R), and when integrated yields the function $\phi(x)$ except as regards an additive constant.

344. *If $f(x)$ be continuous in the interval (a, b), and $F(x)$ denote the indefinite integral $\int_a^x f(x)\,dx$, then, at every point in (a, b), $F(x)$ possesses a differential coefficient which is equal to $f(x)$.*

For since $f(x)$ is continuous, an interval $(x - h_1, x + h_1)$ can be found such that $|f(x \pm \theta h_1) - f(x)| < \epsilon$, for all values of θ numerically less than 1. It follows that

$$F(x \pm h) - F(x) \equiv \int_x^{x\pm h} f(x)\,dx$$

lies between $\pm h\,[f(x) + \epsilon]$ and $\pm h\,[f(x) - \epsilon]$, provided $h < h_1$. Hence, since

$$\frac{F(x \pm h) - F(x)}{\pm h}$$

lies between $f(x) + \epsilon$ and $f(x) - \epsilon$,

for $h < h_1$, it follows that $f(x)$ is the differential coefficient of $F(x)$. At the points a, b, the function $F(x)$ possesses derivatives on the right and on the left respectively, and their values are $f(a)$, $f(b)$.

If $\phi(x)$ be a function which at every point of (a, b) has a differential coefficient, which is a continuous function $f(x)$, then

$$\phi(x) - \phi(a) = \int_a^x f(x)\,dx.$$

For let $\int_a^x f(x)\,dx$ be denoted by $F(x)$, then the function $\phi(x) - F(x)$ has at every point a differential coefficient which is zero, and therefore, by the theorem of § 267, the function $\phi(x) - F(x)$ is constant; it is clear that this constant must be $\phi(a)$; and thus the theorem is established. In this theorem and elsewhere, a derivative at a on the right, and a derivative at b on the left, are included in the term "differential coefficient."

345. If a given bounded function $f(x)$ that is integrable (R) be not everywhere continuous in the interval (a, b), the proof given above is applicable to prove that, *at any point of continuity of $f(x)$, the function $\int_a^x f(x)\,dx$ has a differential coefficient equal to $f(x)$, and thus has a differential coefficient almost everywhere in (a, b).*

That $F(x)$ has a differential coefficient almost everywhere has been shewn, in § 298, to follow from the fact that $F(x)$ is of bounded variation. It has here been proved independently, for the special case of an indefinite integral.

At a point of ordinary discontinuity of $f(x)$, the same proof, when modified by taking only positive values of h, or only negative values of h, and using $f(x+0)$ or $f(x-0)$, in the two cases, instead of $f(x)$, will shew that $F(x)$ has at such a point derivatives on the right and on the left, and that these are $f(x+0)$, $f(x-0)$ respectively. At a point at which $f(x)$ has a discontinuity of the second kind, the proof fails altogether; at such a point therefore $F(x)$ need not possess a differential coefficient, nor definite derivatives on the right and on the left, but may have all its four derivatives $D^+F(x)$, $D_+F(x)$, $D^-F(x)$, $D_-F(x)$ of different values.

If $f(x)$ be integrable (R), and consequently point-wise discontinuous, and $\psi(x)$ be the function formed by extension of the functional values of $f(x)$ at its points of continuity, as explained in § 342, we have

$$f(x) = \chi(x) + \psi(x),$$

where $\chi(x)$ is an integrable null-function; and therefore $f(x)$ and $\psi(x)$ have the same indefinite integral $F(x)$. The derivatives of $F(x)$ are independent of the function $\chi(x)$, and depend only upon $\psi(x)$, which is determined by the values of $f(x)$ at its points of continuity.

Since $F(x+h) - F(x) = \int_x^{x+h} \psi(x)\,dx$, and since the values of $\psi(x)$ in the interval $(x, x+h)$ all lie between $\overline{\psi(x+0)} + \epsilon_1$ and $\underline{\psi(x-0)} - \epsilon_2$, where ϵ_1, ϵ_2 converge to zero as h does so, we see that $\frac{F(x-h)-F(x)}{h}$ lies between $\overline{\psi(x+0)} + \epsilon_1$ and $\underline{\psi(x+0)} - \epsilon_2$, hence $D^-F(x)$, $D_-F(x)$ both lie between[*] $\overline{\psi(x+0)}$ and $\underline{\psi(x+0)}$. By taking h negative, we see that $D^-F(x)$, $D_-F(x)$ both lie between $\overline{\psi(x-0)}$ and $\underline{\psi(x-0)}$. In case $\psi(x)$ have a unique limit on the right, $F(x)$ has a derivative on the right, $\psi(x+0)$; and in case $\psi(x)$ have a unique limit on the left, $F(x)$ has a derivative $\psi(x-0)$ on the left. It may happen that $\psi(x)$ is continuous at a point of discontinuity of $f(x)$; at such a point $F(x)$ has a differential coefficient equal to the value of $\psi(x)$. Even when $\psi(x)$ has a discontinuity of the second kind, it is possible that $F(x)$ may have a differential coefficient, or a derivative on the right or on the left, or both. This case has been considered[†] by G. Prasad.

If $f(x)$ be integrable (R), and $F(x)$ be the corresponding indefinite integral, any one of the four derivatives $DF(x)$, of $F(x)$, is integrable (R), and has $F(x)$ for its indefinite integral.

For $DF(x)$ differs from $f(x)$ only at a point of discontinuity of $f(x)$, and at such a point $DF(x)$ lies between the upper and lower boundaries of $\psi(x)$; thus $f(x) - DF(x)$ is an integrable null-function. Therefore

$$\int_a^x \psi(x)\,dx = \int_a^x D^+F(x)\,dx = \int_a^x D_+F(x)\,dx = \int_a^x D^-F(x)\,dx$$

$$= \int_a^x D_-F(x)\,dx = \int_a^x f(x)\,dx = F(x).$$

It has been shewn that the indefinite integral of a function that is integrable (R) has a differential coefficient at the everywhere dense set of points of continuity of the discontinuous function; there may however also be an everywhere dense set of points at which this continuous function does not possess a differential coefficient.

---

[*] It is stated by Schoenflies, see *Bericht über die Mengenlehre*, p. 208, that the derivatives of $F(x)$ are equal to $\overline{\psi(x+0)}$, $\underline{\psi(x+0)}$, $\overline{\psi(x-0)}$, $\underline{\psi(x-0)}$. This, however, is not necessarily the case. It has been shewn by Hahn, *Monatshefte der Math. u. Physik*, vol. XVI (1905), p. 317, that $f(x)$ may be so chosen that the corresponding integral function has, at a particular point, derivatives on the right having arbitrarily given values lying between, or equal to, the values of $\overline{\psi(x+0)}$, $\underline{\psi(x+0)}$ at the point; and in particular that $f(x)$ may be so constructed as to have, at the point, a definite derivative which has an assigned value between the two limits.

[†] *Bulletin Calcutta Math. Soc.* vol. XV (1924), and vol. XVI (1925).

346. It has been shewn that, if the continuous function $\phi(x)$ possesses everywhere a differential coefficient $f(x)$ which is everywhere a continuous function, then

$$\phi(x) - \phi(a) = \int_a^x f(x)\,dx = F(x).$$

This is a particular case of the following more general theorem:

If $\phi(x)$ be a function continuous in the interval (a, b), and if one of its four derivatives $D^+\phi(x)$, $D_+\phi(x)$, $D^-\phi(x)$, $D_-\phi(x)$ be a bounded R-integrable function in (a, b), then each of the other three derivatives is also bounded and integrable (R) in (a, b), and $\phi(x) - \phi(a)$ is the integral of any one of the four derivatives through the interval (a, x).

If (a, x) be divided into a number of parts (a, x_1), (x_1, x_2), ... (x_{n-1}, x), it has been shewn in § 280, that $\dfrac{\phi(x_r) - \phi(x_{r-1})}{x_r - x_{r-1}}$ lies between the upper and lower boundaries of any one of the four derivatives $D\phi(x)$ in the interval (x_{r-1}, x_r). It follows that $\phi(x) - \phi(a)$ lies between two sums

$$(x_1 - a)\,U(a, x_1, D) + (x_2 - x_1)\,U(x_1, x_2, D) + \ldots + (x - x_{n-1})\,U(x_{n-1}, x, D),$$

$$(x_1 - a)\,L(a, x_1, D) + (x_2 - x_1)\,L(x_1, x_2, D) + \ldots + (x - x_{n-1})\,L(x_{n-1}, x.D)$$

where $U(x_{r-1}, x_r, D)$, $L(x_{r-1}, x_r, D)$ are the upper and lower boundaries of $D\phi(x)$ in the interval (x_{r-1}, x_r); and it is known that these are the same for all four derivatives. The limits of the above sums, when the intervals are diminished indefinitely, so that the greatest of them converges to zero, are the upper and lower integrals of any one of the four functions $D\phi(x)$. If it be known that any one of these derivatives is integrable (R) in (a, x), then the upper and lower integrals are equal, and the other three are also integrable, the common value of the integral being $\phi(x) - \phi(a)$. Thus

$$\phi(x) - \phi(a) = \int_a^x D^+\phi(x)\,dx = \int_a^x D_+\phi(x)\,dx = \int_a^x D^-\phi(x)\,dx = \int_a^x D_-\phi(x)\,dx.$$

It should be observed that, as has been shewn in § 280, the four derivatives are all equal to one another at a point at which one of them is a continuous function; and thus at such a point there is a differential coefficient. If one of the derivatives be integrable, there is therefore a set of points of measure equal to that of the interval (a, b), at which all four derivatives have equal values, and at which therefore a differential coefficient exists.

347. In case $D\phi(x)$ be a bounded function which is not integrable (R), the above proof shews that $\phi(x) - \phi(a)$ lies between the upper and lower integrals, in (a, x), of any one of the four functions $D\phi(x)$. This includes the case in which $\phi(x)$ has a differential coefficient which is bounded but not integrable (R); in that case $\phi(x) - \phi(a)$ lies between

$$\overline{\int_a^x} \phi'(x)\, dx \text{ and } \underline{\int_a^x} \phi'(x)\, dx.$$

Since $\overline{\int_x^{x+h}} \phi'(x)\, dx$ is in absolute value less than hU, where U is the upper boundary of $|\phi'(x)|$ in (a, b), it follows, as in § 343, that $\overline{\int_a^x} \phi'(x)\, dx$ is a continuous function of x; similarly it may be seen that $\underline{\int_a^x} \phi'(x)\, dx$ is a continuous function of x. At a point of continuity of $\phi'(x)$, both

$$\overline{\int_a^x} \phi'(x)\, dx, \quad \underline{\int_a^x} \phi'(x)\, dx$$

have the differential coefficient $\phi'(x)$, as may be seen by a process precisely similar to that in § 344. Thus the upper and lower integrals of $\phi'(x)$ possess properties similar to those of the integral of $\phi'(x)$, when it exists, and both of them may be regarded as primitives of $\phi'(x)$, in the extended sense.

The function $D\phi(x)$, when not integrable (R), may be a non-integrable point-wise discontinuous function, or it may be totally discontinuous.

If $f(x)$ be any non-integrable bounded function, the following theorems may be established by proofs similar to those in § 343 and § 345:

The upper and lower integrals $\overline{\int_a^x} f(x)\, dx$, $\underline{\int_a^x} f(x)\, dx$ *are continuous, and of bounded variation in* (a, b).

At any point of (a, b), *at which* $f(x)$ *is continuous, the upper and lower integrals* $\overline{\int_a^x} f(x)\, dx$, $\underline{\int_a^x} f(x)\, dx$ *each possess a differential coefficient which is equal to* $f(x)$.

348. An important general class of continuous functions for which the four derivatives are not integrable (R), even when a differential coefficient exists, or when derivatives on the right and on the left always exist, is the class of everywhere-oscillating functions. Those functions which become everywhere-oscillating functions when a linear function is added have the same property.

If a derivative $DF(x)$ be such that, in every interval, it has no finite upper boundary or no finite lower boundary, it is certainly not integrable (R); it is therefore only necessary to consider an interval in which the function $DF(x)$ is bounded. Let (a, x) be such an interval, and let us suppose that $F(x) - F(a)$ is not zero.

In every interval (x_{r-1}, x_r) contained in (a, x), $U(x_{r-1}, x_r, D)$ the upper boundary of $DF(x)$ is positive, and $L(x_{r-1}, x_r, D)$ the lower boundary of $DF(x)$ is negative; thus the two sums

$$(x_1 - a)\,U(a, x_1, D) + (x_2 - x_1)\,U(x_1, x_2, D) + \ldots + (x - x_{n-1})\,U(x_{n-1}, x, D),$$
$$(x_1 - a)\,L(a, x_1, D) + (x_2 - x_1)\,L(x_1, x_2, D) + \ldots + (x - x_{n-1})\,L(x_{n-1}, x, D),$$

are such that the first is essentially positive, and the second essentially negative, the non-vanishing number $F(x) - F(a)$ lying between them.

It follows that the limits of these two sums, as the number of subdivisions of (a, x) is increased indefinitely, must be different from one another, since they cannot have zero as their common value; thus

$$\overline{\int_a^x} DF(x)\,dx, \quad \underline{\int_a^x} DF(x)\,dx$$

are distinct from one another.

It has thus been proved that *a continuous function which is everywhere-oscillating in (a, b) cannot have a derivative which is integrable (R) in (a, b), even if it have everywhere a differential coefficient, or definite derivatives on the right and on the left.*

The function $DF(x)$, or $f(x)$, in the case of such function, may be a point-wise discontinuous function such that the measure of the set of points of discontinuity is greater than zero, or it may be a totally discontinuous function.

A continuous monotone function, which is not reducible to a function with an infinite number of oscillations by the addition of a linear function, has at every point definite derivatives on the right and on the left, each of which is either continuous or is an R-integrable discontinuous function, since either derivative has only ordinary discontinuities. Thus such a function has R-integrable derivatives, provided these derivatives are bounded in the interval.

In case the continuous function $F(x)$ have a differential coefficient, or a derivative which is not everywhere finite, or is not bounded in the interval, this derivative is not integrable in the sense in which we have hitherto defined integration. This case will be considered in connection with the theory of improper integrals.

349. The preceding investigations provide answers to the questions which arise as regards the validity of the two propositions (A) and (B) of § 343, which together constitute the fundamental theorem of the Integral Calculus, asserting that the operations of differentiation and of integration are in general reversible. The definition of a definite integral has hitherto been restricted to that of Riemann, and is applicable to bounded functions only. The extensions of that definition to the case of unbounded functions, which will be considered later, and also the more general definition of integration due to Lebesgue, to be considered in Chap. VII, will lead to corresponding extensions of the scope of the fundamental theorem.

As regards the theorem (A), that the indefinite integral

$$F(x) \equiv \int_a^x f(x)\,dx$$

of a bounded R-integrable function possesses a differential coefficient equal, at a point x of (a, b), to $f(x)$, it has been shewn that the theorem holds without restriction in case $f(x)$ is a continuous function; but that, if $f(x)$ be not continuous, the theorem still holds as regards every point of continuity of $f(x)$. It follows that the points of (a, b) at which $F(x)$ either possesses no differential coefficient, or possesses one which is not equal to $f(x)$, form a set of measure zero, which may however be everywhere dense in (a, b).

The theorem (B) that, if $\phi(x)$ possess a differential coefficient $f(x)$, then the corresponding indefinite integral $F(x) \equiv \int_a^x f(x)\,dx$ differs from $\phi(x)$ only by a constant, holds if $f(x)$ be a continuous function, and more generally, if $f(x)$ be bounded and integrable (R). In case $\phi(x)$ does not at all points possess a differential coefficient, the more general theorem is applicable that, if any one of the four derivatives of $\phi(x)$ be bounded and integrable (R), then the integral function corresponding to that derivative differs from $\phi(x)$ by a constant only. The theorem fails either in case $\phi(x)$ be not a function with bounded derivatives, or in case it be a function with bounded derivatives, but those derivatives do not satisfy Riemann's condition of integrability.

The problem of the determination of a continuous function which shall have a given function $f(x)$ for its differential coefficient, at every point at which $f(x)$ is continuous, may be considered here in the case in which $f(x)$ is restricted to be bounded in the interval (a, b) for which it is defined. This problem of the determination of such a primitive of $f(x)$ is regarded as having a unique solution provided functions exist which satisfy the condition, and further provided any two such functions differ from one another by a constant only, that constant having one and the same value for the whole interval (a, b). In the first place, the problem cannot be determinate unless $f(x)$ be integrable (R); for either of the two functions $\overline{\int_a^x} f(x)\,dx$, $\underline{\int_a^x} f(x)\,dx$ satisfies the condition of the problem, and these functions do not differ from one another by a constant, as they both vanish at the point a, and are elsewhere unequal. Next, if $f(x)$ be integrable (R), the function $\int_a^x f(x)\,dx$ satisfies the condition of the problem, but the solution is not necessarily unique. In case however the points of discontinuity of the R-integrable function $f(x)$ form an enumerable set, the theorem of § 267 shews that the solution is determinate; for any two functions which have equal finite differential coefficients at all points of (a, b) except those of an enumerable set, differ from one another by a constant. In this case $\int_a^x f(x)\,dx + C$ is the function required. When the points of discontinuity of the integrable function $f(x)$ form an unenumerable set, with a perfect nucleus, although that set must have zero measure, the problem has not a unique solution. For, although $\int_a^x f(x)\,dx$ is a function which has the required property, another solution is obtained by adding to it any continuous function which has all the intervals complementary to the perfect component of the unenumerable set as lines of invariability; that such functions exist has been established in § 269. There exists however only one function, viz. $\int_a^x f(x)\,dx$, with bounded derivatives, which satisfies the condition of the problem; for it has been shewn in § 286, that any two functions which have bounded derivatives, one of which derivatives is prescribed at all points not belonging to a certain set of measure zero, differ from one another by a constant.

Similar remarks apply to the more general problem of the determination of a function which shall have one of its four derivatives, say the upper one on the right, equal to a given function $f(x)$ at every point of continuity of $f(x)$. This problem has a solution whenever $f(x)$ is bounded; in virtue of the theorem of § 267, the solution is unique when $f(x)$ is integrable (R), and the points of discontinuity of $f(x)$ form an enumerable set. When $f(x)$ is integrable (R), and the set of points of discontinuity is unenumerable, there exists, in virtue of the theorem of § 286, only one solution for which the derivatives are bounded. As before, if the restriction, that the required function is to have bounded derivatives, be not imposed, the solution of the problem is indeterminate.

EXAMPLE

The following example was given* by Volterra, as the first case of a continuous function possessing a bounded differential coefficient, not integrable (R).

Let G be a perfect non-dense set of points in the interval (a, b), and such that its content is greater than zero. Let (α, β) be an interval complementary to the set G, and let $\phi(x, \alpha) = (x - \alpha)^2 \sin \frac{1}{x - \alpha}$, and therefore $\phi'(x, \alpha) = 2(x - \alpha) \sin \frac{1}{x - \alpha} - \cos \frac{1}{x - \alpha}$. The function $\phi'(x, \alpha)$ vanishes at an infinite number of points in (α, β); let $\alpha + \gamma$ be the greatest value of x which does not exceed $\frac{1}{2}(\alpha + \beta)$, for which $\phi'(x, \alpha)$ vanishes. Let $F(x) = 0$ at every point of G, and in each interval (α, β) complementary to G, let $F(x) = \phi(x, \alpha)$, for values of x such that $\alpha \leqq x \leqq \alpha + \gamma$; let $F(x) = \phi(\alpha + \gamma, \alpha)$, for values of x such that $\alpha + \gamma \leq x \leqq \beta - \gamma$; and let $F(x) = -\phi(x, \beta)$, for $\beta - \gamma \leqq x \leqq \beta$. The function $F(x)$ is continuous, and has everywhere a finite differential coefficient which is bounded in the interval (a, b). It is easily seen that $F'(x)$ vanishes at every point of G. The function $F'(x)$ has a discontinuity of measure > 2 at each point of the set G, which is not of zero content, and therefore $F'(x)$ is not an integrable function.

---

* *Giorn. di Battaglini*, vol. XIX (1881), p. 335.

FUNCTIONS WHICH ARE LINEAR IN EACH INTERVAL OF A SET

350. The existence of continuous functions which are linear in each interval of an everywhere dense set of intervals has been already referred to in § 274. It has been shewn in § 269 how a function $f(x)$ can be constructed which is continuous, and has as lines of invariability the intervals complementary to a non-dense perfect set of points. It is clear that the integral function $\int_a^x f(x)\,dx$ is linear in each of the intervals, and being also continuous, it is a function of the type referred to. A more general function which is continuous, and is linear in each interval of the set, may be obtained by adding to $\int_a^x f(x)\,dx$ any continuous function for which the intervals of the set are lines of invariability.

INTEGRATION BY PARTS

351. If u, v denote functions of x, defined for the interval (a, b), and which have continuous differential coefficients $\frac{du}{dx}, \frac{dv}{dx}$ in that interval, the formula

$$\frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$$

leads at once, by integration over the interval (a, b), to the well-known formula

$$\int_a^b u\frac{dv}{dx}\,dx + \int_a^b v\frac{du}{dx}\,dx = \Big[uv\Big]_a^b$$

for integration by parts, due to Leibnitz.

The simplest generalization* of this formula is that, if $f(x)$, $g(x)$ are bounded and integrable (R) in the interval (a, b), then

$$\int_a^b f(x)\left\{\int_a^x g(x)\,dx\right\}dx + \int_a^b g(x)\left\{\int_a^x f(x)\,dx\right\}dx = \int_a^b f(x)\,dx\int_a^b g(x)\,dx.$$

To prove this formula, we employ the identity

$$\sum_{r=1}^{r=p}\alpha_r\sum_{r=1}^{r=p}\beta_r = \sum_{r=1}^{r=p}\alpha_r s'_{r-1} + \sum_{r=1}^{r=p}\beta_r s_r,$$

where s_r denotes $\alpha_1 + \alpha_2 + \ldots + \alpha_r$, and s'_{r-1} denotes $\beta_1 + \beta_2 + \ldots + \beta_{r-1}$.

---

* See Thomae, *Zeitschr. f. Math.* vol. XX (1875), pp. 475–478; also Hardy, *Messenger of Math.* vol. XXX (1900), pp. 185–187, and vol. XLVIII (1919), p. 90.

In a net D_n fitted on to the interval (a, b), let $\alpha_r = \delta_r M_1(\delta_r)$, where $M_1(\delta_r)$ is a number not greater than the upper boundary of $f(x)$ in the mesh δ_r, nor less than the lower boundary of $f(x)$ in δ_r; let β_r be $\delta_r M_2(\delta_r)$, where $M_2(\delta_r)$ is defined similarly in relation to $g(x)$. We have then, if $p = m_n$,

$$\sum_{r=1}^{r=m_n}\left[\delta_r M_1(\delta_r)\sum_{t=1}^{t=r-1}\delta_t M_2(\delta_t)\right] + \sum_{r=1}^{r=m_n}\left[\delta_r M_2(\delta_r)\sum_{t=1}^{t=r}\delta_t M_1(\delta_t)\right]$$
$$= \sum_{r=1}^{r=m_n}\delta_r M_1(\delta_r)\sum_{r=1}^{r=m_n}\delta_r M_2(\delta_r) \;\ldots\ldots\ldots(1).$$

Let n be so great that $\int_a^b f(x)\,dx$ differs from both $\sum_{r=1}^{r=m_n}\delta_r U_1(\delta_r)$ and $\sum_{r=1}^{r=m_n}\delta_r L_1(\delta_r)$ by less than η: and so great that $\int_a^b g(x)\,dx$ differs from both $\sum_{r=1}^{r=m_n}\delta_r U_2(\delta_r)$ and $\sum_{r=1}^{r=m_n}\delta_r L_2(\delta_r)$ by less than η; where η is an arbitrarily chosen positive number. $U_1(\delta_r)$, $L_1(\delta_r)$ are the upper and lower boundaries of $f(x)$ in the mesh δ_r, and $U_2(\delta_r)$, $L_2(\delta_r)$ those of $g(x)$.

Then it is clear that, for every value of r,

$$\sum_{t=1}^{t=r}\delta_t U_1(\delta_t) - \int_a^{x_r} f(x)\,dx < \eta, \text{ and } \int_a^{x_r} f(x)\,dx - \sum_{t=1}^{t=r}\delta_t L_1(\delta_t) < \eta,$$

hence $\int_a^{x_r} f(x)\,dx$ and $\sum_{t=1}^{t=r}\delta_t M_1(\delta_t)$ differ from one another by less than η. A similar result holds with respect to the function $g(x)$; the mesh δ_r being (x_{r-1}, x_r).

If $M_1(\delta_r) = f(x_{r-1})$, and $M_2(\delta_r) = g(x_{r-1})$, we see that

$$\sum_{r=1}^{r=m_n}\left[\delta_r M_1(\delta_r)\sum_{t=1}^{t=r-1}\delta_t M_2(\delta_t)\right]$$

differs from $\sum_{r=1}^{r=m_n}\delta_r f(x_{r-1})\int_a^{x_{r-1}} g(x)\,dx$ by less than $\eta\sum_{r=1}^{r=m_n}\delta_r\,|f(x_{r-1})|$ or by less than $A(b-a)\eta$, where A is the upper boundary of $|f(x)|$ in (a, b). A similar result holds with respect to $g(x)$.

As η is arbitrarily small, we now see that, when $n \sim \infty$, the formula obtained by proceeding to the limit of the expressions in (1) gives the above formula for integration by parts.

The formula is equivalent to a somewhat more complex formula which was given* by Du Bois-Reymond.

Let α, β be fixed points in the interval (a, b). Writing

$$F(x) = \int_a^x f(x)\,dx, \quad G(x) = \int_a^x g(x)\,dx,$$

---

* *Abhandlungen d. Munch. Akad.* vol. XII (1875), p. 129.

the above formula becomes

$$\int_a^b f(x)\,G(x)\,dx + \int_a^b g(x)\,F(x)\,dx = \int_a^b f(x)\,dx \int_a^b g(x)\,dx.$$

If we subtract from each side the expression

$$G(\beta)\int_a^b f(x)\,dx + F(\alpha)\int_a^b g(x)\,dx,$$

and express each of the integrals on the right-hand side as the sum of two integrals, taken from a to α and from α to b, in the first case, and from a to β and β to b, in the second case, we find, after a little reduction, Du Bois-Reymond's formula for integration by parts:

$$\int_a^b f(x)\left\{\int_\beta^x g(x)\,dx\right\}dx + \int_a^b g(x)\left\{\int_\alpha^x f(x)\,dx\right\}dx$$
$$= \left[\int_\alpha^x f(x)\,dx \int_\beta^x g(x)\,dx\right]_a^b.$$

This formula reduces to the earlier one, if we take $\alpha = \beta = a$.

The case of functions $f(x^{(1)}, x^{(2)})$, $g(x^{(1)}, x^{(2)})$, both integrable (R) in a cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, may be considered by applying the formula (1), when the meshes δ are those of a net of a system fitted on to the cell. The form of the resulting expression when we proceed to the limit will depend upon the order in which the meshes of the net D_n are arranged in the series. If we count the meshes from left to right, first taking the lowest row along the $x^{(1)}$ axis, then the next row also from left to right, and so on, it is easily seen that the formula obtained is

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)}) \left[\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, x^{(2)})} g(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)})\right] d(x^{(1)}, x^{(2)})$$
$$+ \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} g(x^{(1)}, x^{(2)}) \left[\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, x^{(2)})} f(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)})\right] d(x^{(1)}, x^{(2)})$$
$$= \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)}) \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} g(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)}).$$

This is one of the formulae analogous to the formula for integration by parts in the case of functions of one variable.

CAUCHY'S DEFINITION OF AN IMPROPER INTEGRAL

352. The definition of the R-integral of a function $f(x)$, in a linear interval (a, b), presupposes that $f(x)$ is bounded in (a, b). A method was given by Cauchy of extending the definition to cases in which $f(x)$ is unbounded in the neighbourhoods of points of a finite set.

Let us suppose that a point c, where $a < c < b$, is such that, in its arbitrarily small neighbourhood, the absolute values of the function have no upper limit, and let us suppose further that c is the only point of this kind, and that the function $f(x)$ is integrable in any sub-interval of (a, b) which does not contain c within it, or at an end. The two integrals

$$\int_a^{c-\epsilon} f(x)\,dx,\quad \int_{c+\epsilon'}^b f(x)\,dx$$

both exist, whatever sufficiently small positive values be assigned to ϵ, ϵ'.

It may happen that, as ϵ, ϵ' *are diminished independently, so as to converge in each case to the limit zero, the two integrals also converge to definite limits; if this be the case, we define the sum*

$$\lim_{\epsilon\sim 0}\int_a^{c-\epsilon} f(x)\,dx + \lim_{\epsilon'\sim 0}\int_{c+\epsilon'}^b f(x)\,dx$$

to be the improper integral of $f(x)$ *in the interval* (a, b), *and we denote this improper integral by*

$$\int_a^b f(x)\,dx,$$

using the same notation as in the case in which $f(x)$ *is integrable* (R) *in* (a, b).

The condition that $\displaystyle\lim_{\epsilon\sim 0}\int_a^{c-\epsilon} f(x)\,dx$

should exist, is that, corresponding to each arbitrarily small number δ which may be chosen, a number ϵ_1 can be found, such that

$$\left|\int_{c-\theta\epsilon_1}^{c-\epsilon_1} f(x)\,dx\right| < \delta,$$

whatever value θ may have, subject to the condition $0 < \theta < 1$.

A similar condition must be satisfied in order that

$$\lim_{\epsilon'\sim 0}\int_{c+\epsilon'}^b f(x)\,dx$$

may exist.

It may happen that, although the two limits

$$\int_a^{c-\epsilon} f(x)\,dx,\quad \int_{c+\epsilon'}^b f(x)\,dx$$

do not exist, yet if we take $\epsilon' = \epsilon$, the sum

$$\int_a^{c-\epsilon} f(x)\,dx + \int_{c+\epsilon}^b f(x)\,dx$$

may have a definite limit; when this is the case, the limit defines Cauchy's *principal value* of the integral of $f(x)$, in (a, b).

It thus appears that a principal value may exist when the function possesses neither an integral nor an improper integral in the interval (a, b).

In case the point a *itself be a point of infinite discontinuity, then the limit*

$$\int_{a+\epsilon}^b f(x)\,dx,$$

for $\epsilon \sim 0$, when it exists, is defined to be the improper integral

$$\int_a^b f(x)\,dx$$

of $f(x)$ in the interval (a, b).

A similar definition applies in case the point b is a point of infinite discontinuity of the function.

If, in the interval (a, b), there are two points of infinite discontinuity c_1, c_2 (where $a < c_1 < c_2 < b$), let c be any point between c_1 and c_2. Then in case the four improper integrals

$$\int_a^{c_1} f(x)\,dx, \quad \int_{c_1}^{c} f(x)\,dx, \quad \int_c^{c_2} f(x)\,dx, \quad \int_{c_2}^{b} f(x)\,dx$$

all exist, their sum is defined to be the improper integral of $f(x)$ in (a, b), and is denoted by

$$\int_a^b f(x)\,dx;$$

and it is clearly independent of the value of c. The definition, in case one of the points c_1, c_2 is an end-point of the interval (a, b), is of the same character; or both of them may be end-points. If $c_1 = a$, $c_2 = b$, then if the two improper integrals

$$\int_a^c f(x)\,dx, \quad \int_c^b f(x)\,dx$$

exist, their sum defines the improper integral

$$\int_a^b f(x)\,dx.$$

The definition of an improper integral can now immediately be extended to the case in which there is any finite number of points of infinite discontinuity in the interval. If these be $c_1, c_2, c_3, \ldots c_n$, taken in order from left to right, and the improper integrals

$$\int_a^{c_1} f(x)\,dx, \quad \int_{c_1}^{c_2} f(x)\,dx, \ldots \int_{c_n}^{b} f(x)\,dx$$

all exist, their sum is defined to be the improper integral

$$\int_a^b f(x)\,dx.$$

The definition can be extended to the case in which there is an infinite number of points of infinite discontinuity of $f(x)$, forming a set G, of the first species, and of the first order. In that case G' consists of a finite number of points $c_1, c_2, \ldots c_n$.

If the improper integrals

$$\int_a^{c_1-\epsilon_1} f(x)\,dx, \quad \int_{c_1+\epsilon_1'}^{c_2-\epsilon_2} f(x)\,dx, \quad \int_{c_2+\epsilon_2'}^{c_3-\epsilon_3} f(x)\,dx, \ldots,$$

each of which falls under the last case, all exist, and have values which converge to definite numbers as $\epsilon_1, \epsilon_1', \epsilon_2, \epsilon_2', \ldots$ converge to zero, independently of one another, the sum of their limits is taken to define the improper integral

$$\int_a^b f(x)\,dx.$$

It is clear that this definition admits of extension to the case in which G is of the first species, and of any order.

It is easily seen that an integral which exists in accordance with this definition is such that the integral also exists in the interval (a, x), where x is any point in (a, b), and that it is a continuous function of x. Moreover

$$\int_a^b f(x)\,dx = \int_a^x f(x)\,dx + \int_x^b f(x)\,dx.$$

This definition of an improper integral was given* by Du Bois-Reymond, for the case in which G' is finite. The definition has been extended† by Schoenflies to the case in which G is enumerable, but possesses transfinite derivatives.

353. When the integral of a function $f(x)$, with a finite, or an enumerable, set of points of infinite discontinuity, exists, it may, or may not, be the case that $|f(x)|$ has an integral in the same interval.

The improper integral $\int_a^b f(x)\,dx$

is said to be absolutely convergent in the interval (a, b), *in case*

$$\int_a^b |f(x)|\,dx$$

exists; otherwise it is said to be conditionally, or non-absolutely, convergent.

The integral is said to be absolutely convergent at a point c, of infinite discontinuity of the function, when both the integrals

$$\int_{c-\epsilon}^{c} |f(x)|\,dx, \quad \int_c^{c+\epsilon} |f(x)|\,dx$$

exist, provided ϵ be sufficiently small. If either of these does not so exist, the integral

$$\int_a^b f(x)\,dx$$

is non-absolutely convergent at the point c.

An absolutely convergent integral is absolutely convergent at each of the points of infinite discontinuity of the function. The converse also holds.

---

[*] *Crelle's Journal*, vol. LXXIX (1875), p. 36.

[†] See *Bericht*, vol. I, p. 185.

If $\int_a^b |f(x)|\,dx$ exists, then $\int_a^b f(x)\,dx$ exists. For, if c be a point of infinite discontinuity of $f(x)$, we have

$$\left|\int_{c+\epsilon'}^{c+\epsilon} f(x)\,dx\right| \leqq \int_{c+\epsilon'}^{c+\epsilon} |f(x)|\,dx < \eta;$$

for all sufficiently small values of ϵ, whatever value ϵ' ($< \epsilon$) may have. The corresponding condition, on the other side of c, is also seen to be satisfied. The condition of convergence of $\int_a^b f(x)\,dx$, at the point c, is therefore satisfied.

EXAMPLES

1. The integral $\int_0^1 \frac{1}{x}\sin\frac{1}{x}\,dx$ exists, but is non-absolutely convergent. For consider $\int \left|\frac{1}{x}\sin\frac{1}{x}\right| dx$, taken over the set of intervals

$$\left(\frac{1}{2n\pi + \frac{1}{2}\pi}, \frac{1}{2n\pi + \frac{1}{4}\pi}\right), \text{ where } n = 1, 2, 3, \ldots m.$$

We have
$$\int_{\frac{1}{2n\pi+\frac{1}{2}\pi}}^{\frac{1}{2n\pi+\frac{1}{4}\pi}} \left|\frac{1}{x}\sin\frac{1}{x}\right| dx > \frac{1}{\sqrt{2}}\log_e\left(1 + \frac{1}{8n - 1}\right),$$

hence the integral taken over the set of intervals is

$$> \frac{1}{\sqrt{2}}\log_e \prod_{n=1}^{n=m}\left(1 + \frac{1}{8n+1}\right) > \frac{1}{\sqrt{2}}\log_e\left\{\sum_{n=1}^{n=m}\frac{1}{8n+1}\right\};$$

thus, as m increases indefinitely, the value of the integral does so, and therefore $\frac{1}{x}\sin\frac{1}{x}$ cannot be integrable in the interval (0, 1), in accordance with Cauchy's definition. On the other hand $\int_\epsilon^1 \frac{1}{x}\sin\frac{1}{x}\,dx$ converges to a finite limit, as $\epsilon \sim 0$. so that the integral $\int_0^1 \frac{1}{x}\sin\frac{1}{x}\,dx$ exists in accordance with Cauchy's definition, but it is not absolutely convergent.

2. Let $f(x)$ denote a function which is integrable (R) in every interval (a, b), where $0 < a < b$; and let $f(x)$ be, in the neighbourhood of the point 0, of the form $\frac{\phi(x)}{x^k}$, where k is positive, and $\phi(x)$ is a bounded function, defined for the interval $(0, c)$.

We have
$$\left|\int_\epsilon^{\epsilon'} \frac{\phi(x)}{x^k}\,dx\right| < A\int_\epsilon^{\epsilon'} \frac{dx}{x^k} < \frac{A}{1-k}\left[\epsilon'^{1-k} - \epsilon^{1-k}\right],$$

where A is some positive number.

If $0 < k < 1$, it is clear that $\int_\epsilon^{\epsilon'} \frac{\phi(x)}{x^k}\,dx$ is arbitrarily small, for a sufficiently small value of ϵ' ($> \epsilon$), and therefore the improper integral $\int_0^c f(x)\,dx$ exists, being convergent at the point $x = 0$. If $k \geqq 1$, the improper integral does not exist.

3. Let $f(x)$ be, in the neighbourhood on the right of the point 0, of the form $\frac{\phi(x)}{x[\log x]^{1+p}}$, where p is positive, and $\phi(x)$ satisfies the same condition as in Ex. 2.

We have
$$\left|\int_\epsilon^{\epsilon'} \frac{\phi(x)}{x[\log x]^{1+p}}\,dx\right| < \frac{A}{p}\left\{[\log\epsilon]^{-p} - [\log\epsilon']^{-p}\right\};$$

and thus the improper integral $\int_0^c f(x)\,dx$ exists, being absolutely convergent.

4. $\int \frac{\tan x}{x} dx$, taken through any interval which contains a point of infinite discontinuity of $\tan x$, does not exist.

For
$$\int_{\frac{\pi}{2}-\epsilon}^{\frac{\pi}{2}-\epsilon'} \frac{\tan x}{x} dx > \frac{2}{\pi} \log \frac{\sin \epsilon}{\sin \epsilon'},$$
and this is arbitrarily great, for a sufficiently great value of ϵ/ϵ'; thus the integral does not converge at the point $x = \frac{1}{2}\pi$. The integral possesses however a principal value at the point $\frac{1}{2}\pi$. For the sum of the integrals taken through the intervals $(\frac{1}{2}\pi - \epsilon, \frac{1}{2}\pi - \epsilon')$ and $(\frac{1}{2}\pi + \epsilon', \frac{1}{2}\pi + \epsilon)$ is
$$\int_{\epsilon'}^{\epsilon} \cot x \,.\, \frac{2x}{\frac{1}{4}\pi^2 - x^2} dx < \pi (\tfrac{1}{4}\pi^2 - \epsilon^2)^{-1} \sin \epsilon,$$
and this converges to 0, with ϵ.

5. The function $\cos (e^{\frac{1}{x}}) + \frac{1}{x} e^{\frac{1}{x}} \sin (e^{\frac{1}{x}})$ oscillates between indefinitely great positive and negative values, in the neighbourhood of the point $x = 0$. For every value of x, except $x = 0$, the function $= \frac{d}{dx} \{x \cos (e^{\frac{1}{x}})\}$.

Also
$$\int_{\epsilon'}^{\epsilon} \frac{d}{dx} \{x \cos (e^{\frac{1}{x}})\} dx = \epsilon \cos (e^{\frac{1}{\epsilon}}) - \epsilon' \cos (e^{\frac{1}{\epsilon'}}),$$
where $\epsilon > \epsilon' > 0$. It thus appears that the integral of the function converges at the point $x = 0$; and therefore the function is integrable in an interval containing that point

RIEMANN INTEGRALS OVER AN UNBOUNDED INTERVAL

354. The definition of the integral of a bounded integrable function given in § 330 is applicable only to the case in which both the limits a, b are definite points, and in which therefore the interval of integration is finite.

Let $x_1, x_2, \dots x_n, \dots$ be a sequence of increasing numbers having no upper limit; it may then happen that the sequence of R-integrals
$$\int_a^{x_1} f(x)\, dx, \quad \int_a^{x_2} f(x)\, dx, \dots \quad \int_a^{x_n} f(x)\, dx, \dots$$
has a definite limit A, independent of the particular sequence $\{x_n\}$ chosen. When this is the case $f(x)$ is said to have an integral (R)
$$\int_a^{\infty} f(x)\, dx,$$
in the unbounded interval (a, ∞), the value of this integral being A. It has been presupposed that, in every interval (a, x), the function $f(x)$ is integrable (R).

If the integrals
$$\int_{x_1}^{b} f(x)\, dx, \quad \int_{x_2}^{b} f(x)\, dx, \dots \quad \int_{x_n}^{b} f(x)\, dx, \dots,$$
where $x_1, x_2, \dots x_n, \dots$ form a sequence of decreasing values of x which has

no lower limit, all exist, and the sequence of integrals have a limit B. independent of the particular sequence chosen, the limit B is denoted by

$$\int_{-\infty}^{b} f(x)\,dx.$$

If the two integrals

$$\int_{0}^{\infty} f(x)\,dx, \quad \int_{-\infty}^{0} f(x)\,dx,$$

as thus defined, both exist, their sum is denoted by

$$\int_{-\infty}^{\infty} f(x)\,dx.$$

The three numbers

$$\int_{a}^{\infty} f(x)\,dx, \quad \int_{-\infty}^{b} f(x)\,dx, \quad \int_{-\infty}^{\infty} f(x)\,dx,$$

being the limits of integrals, and not themselves in the proper sense of the term integrals, belong to the class of improper integrals.

In each case it is necessary, but not sufficient, for the existence of these improper integrals, that $f(x)$ be integrable (R) in every finite interval contained in the intervals (a, ∞), $(-\infty, b)$, or $(-\infty, \infty)$; and it will at present be assumed that $f(x)$ is bounded in every such finite interval, and has therein a proper integral.

In case the integral $\int_{-c}^{c} f(x)\,dx$ have a definite limit, as c is indefinitely increased, that limit is said to be the *principal value* of

$$\int_{-\infty}^{\infty} f(x)\,dx.$$

This principal value may exist, even when the integral

$$\int_{-\infty}^{\infty} f(x)\,dx,$$

as defined above, does not exist; but in case the latter does exist, its value is equal to its principal value.

The necessary and sufficient conditions for the existence of the integral

$$\int_{a}^{\infty} f(x)\,dx,$$

are (1), *that the integral exists in every interval* (a, x), *where* $x > a$, *and* (2), *that, corresponding to every arbitrarily chosen positive number* ϵ, *a value* ξ, *of* x, *can be found such that*

$$\left| \int_{\xi}^{\xi'} f(x)\,dx \right| < \epsilon,$$

for every value of ξ', *such that* $\xi' > \xi$.

A similar condition applies to the case of

$$\int_{-\infty}^{b} f(x)\,dx.$$

355. It was shewn in § 332 that the necessary and sufficient condition that the bounded function $f(x)$ be integrable (R) in the interval (a, b) is that, for a particular system (D_n) of nets, fitted on to (a, b), the sum $\sum_{m=1}^{m=m_n} \delta_m^{(n)} F(\delta_m^{(n)})$ should converge to zero, as $n \sim \infty$. It was also shewn that, if this condition be satisfied for one system of nets, it is satisfied for all such systems.

We have to enquire how far a corresponding condition applies to the case of an R-integral through an infinite interval.

Since $\int_a^\infty f(x)\,dx$, when it exists, is the limit $\lim_{b\sim\infty} \int_a^b f(x)\,dx$, we see that the integral is given as the repeated limit $\lim_{b\sim\infty} \lim_{n\sim\infty} \Sigma\delta_m^{(n)} U(\delta_m^{(n)})$, or as $\lim_{b\sim\infty} \lim_{n\sim\infty} \Sigma\delta_m^{(n)} L(\delta_m^{(n)})$. The question then arises whether, or under what conditions, the order of the repeated limits may be reversed without altering their values. When this can be done, we have

$$\int_a^\infty f(x)\,dx = \lim_{n\sim\infty} \sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)}) = \lim_{n\sim\infty} \sum_{m=1}^{\infty} \delta_m^{(n)} L(\delta_m^{(n)}),$$

where the summations are taken for the meshes of a net fitted on to the infinite interval (a, ∞).

In the first place it is necessary that the two sums should not be divergent.

The following theorem will be established:

If, for the function $f(x)$, bounded in the interval (a, ∞), a system of nets fitted on to the interval exists, such that $\sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)})$, $\sum_{m=1}^{\infty} \delta_m^{(n)} L(\delta_m^{(n)})$ both exist, as definite numbers, $\overline{\Sigma}_n$, $\underline{\Sigma}_n$, which converge, as $n \sim \infty$, to one and the same number A, then $\int_a^\infty f(x)\,dx$ exists, and its value is A.

From the condition stated in the theorem it follows that

$$\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$$

exists, and converges to 0, as $n \sim \infty$. It will be shewn that this condition is sufficient to ensure that $\int_a^b f(x)\,dx$ exists in every finite interval (a, b), where $b > a$. If b is not an end-point of a mesh of the system, we may modify the system of nets by dividing into two each mesh that contains the point b. The first sum will thereby be not increased, and the second sum will not be diminished; accordingly the condition that $\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$ exists and converges to zero, as $n \sim \infty$, will still hold good. It then follows that, when we take the finite sum $\sum_{m=1}^{m=m_n} \delta_m^{(n)} F(\delta_m^{(n)})$, where b is an end-point of $\delta_{m_n}^{(n)}$, this sum will converge to zero, as $n \sim \infty$: and thus $\int_a^b f(x)\,dx$ exists as an R-integral.

Again we have, for any finite value $\bar{m}$, of m,

$$\sum_{m=1}^{m=\bar{m}} \delta_m^{(n)} U(\delta_m^{(n)}) - \int_a^{a+\sum_1^{\bar{m}}\delta_m^{(n)}} f(x)\,dx \leq \sum_{m=1}^{m=\bar{m}} \delta_m^{(n)} F(\delta_m^{(n)});$$

and thus $< \eta$, where η is an arbitrarily chosen positive number, and n is chosen so large that $\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)}) < \eta$. The number $\bar{m}$ may be chosen so large that $\left| \sum_{m=\bar{m}+1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)}) \right| < \eta$; hence

$$\sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)}) - \int_a^{a+\sum_1^{m}\delta_m^{(n)}} f(x)\,dx < 2\eta,$$

for $m > m_1$, some fixed integer.

We may also suppose n to have been chosen so large that

$$\sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)}) - A < \eta;$$

we see then that $\int_a^{a+\xi} f(x)\,dx$ differs from A by less than 3η, for all values of ξ which are equal to $a + \sum_{m=1}^{m} \delta_m^{(n)}$, where $m \geq m_1$. If X be any number greater than the least of these values of ξ, two values of ξ exist, between which X lies. If ξ_1 be the smaller of these, $\int_{\xi_1}^{X} f(x)\,dx$ lies between Ud and Ld, where d is the maximum of the meshes d_n of D_n; and U, L are the upper and lower boundaries of $f(x)$ in (a, ∞). It now follows that $\int_a^X f(x)\,dx$ differs from A by less than $3\eta + \zeta$, where ζ is the greater of the two numbers $|Ud|$, $|Ld|$. Since η, ζ are both arbitrarily small, it follows that $\int_a^X f(x)\,dx$ converges to A, as $X \sim \infty$; and thus that the R-integral $\int_a^{\infty} f(x)\,dx$ exists.

It should be observed that the convergence of $\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$ to zero, as $n \sim \infty$, for a particular system of nets, is not by itself sufficient to ensure the existence of $\int_a^{\infty} f(x)\,dx$, but only that of $\int_a^b f(x)\,dx$, for every finite value of b $(> a)$. In this respect an integral (R) over an infinite interval differs from one over a finite interval; since, in the latter case, the convergence of the finite sum $\Sigma\delta^{(n)}F(\delta^{(n)})$ to zero, as $n \sim \infty$, for a particular system of nets, is sufficient to ensure the existence of the integral.

356. The converse theorem will now be proved, that:

If $\int_a^\infty f(x)\,dx$ *have a definite value, a particular system of nets can always be determined for* (a, ∞), *such that* $\sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)})$ *exists for each net* D_n *of the system, and converges to the value of the integral, as* $n \sim \infty$.

Let us consider a set of points $a, x_1, x_2, \ldots x_n, \ldots$ which diverges as $n \sim \infty$. A net can be determined for the interval (a, x_1), such that

$$\Sigma\delta U(\delta) - \int_a^{x_1} f(x)\,dx < \tfrac{1}{2}\epsilon;$$

similarly, a net can be determined for (x_1, x_2), such that

$$\Sigma\delta U(\delta) - \int_{x_1}^{x_2} f(x)\,dx < \frac{1}{2^2}\epsilon;$$

and generally, a net can be determined for (x_{n-1}, x_n), such that

$$\Sigma\delta U(\delta) - \int_{x_{n-1}}^{x_n} f(x)\,dx < \frac{1}{2^n}\epsilon.$$

Thus a net can be fitted on to (a, x_n), for which

$$\Sigma\delta U(\delta) - \int_a^{x_n} f(x)\,dx < \epsilon\left(1 - \frac{1}{2^n}\right).$$

Now x_n can be taken so large that $\left|\int_a^{x_n} f(x)\,dx - \int_a^{\infty} f(x)\,dx\right| < \eta$; therefore a net can be fitted on to (a, ∞), such that $\sum_{n=1}^{m} \delta U(\delta)$ differs from $\int_a^\infty f(x)\,dx$ by less than an arbitrarily small number, for all sufficiently large values of m. Therefore $\sum_{m=1} \delta U(\delta)$ converges to a limit which differs from $\int_a^\infty f(x)\,dx$ by not more than ϵ. By taking a monotone sequence of values of ϵ which converges to zero, we obtain a system of nets, fitted on to (a, ∞), such as is required.

It may also be possible to define a system of nets D_n such that $\Sigma\delta U(\delta)$, taken over (a, b), does not converge, as $b \sim \infty$; and thus the theorem

$$\int_a^\infty f(x)\,dx = \lim_{n\sim\infty} \sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)}) = \lim_{n\sim\infty} \sum_{n=1}^{\infty} \delta_m^{(n)} L(\delta_m^{(n)})$$

only holds, provided the nets are such that the sums exist for each net of the system, from and after some fixed value of n.

It can also be shewn that:

When the integral $\int_a^\infty f(x)\,dx$ *has a definite value, then for any system of nets fitted on to* (a, ∞) *which is such that* $\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$ *exists, and converges to zero, as* $n \sim \infty$, *the set of numbers* $\sum_{m=1}^{\infty} \delta_m^{(n)} U(\delta_m^{(n)})$ *exists, and converges to the value of the integral.*

Since, for every value of m,

$$\int_a^{a+\overset{m}{\Sigma}\delta_m^{(n)}} f(x)\,dx \le \sum_1^m \delta_m^{(n)} U(\delta_m^{(n)}) \le \int_a^{a+\overset{m}{\Sigma}\delta_m^{(n)}} f(x)\,dx + \sum_1^m \delta_m^{(n)} F(\delta_m^{(n)}),$$

we see that the number $\overset{m}{\Sigma}\delta_m^{(n)} U(\delta_m^{(n)})$ has, as $m \sim \infty$, limits of indeterminancy between $\int_a^\infty f(x)\,dx$ and $\int_a^\infty f(x)\,dx + \sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$. Therefore as $n \sim \infty$, $\overset{\infty}{\Sigma}\delta_m^{(n)} U(\delta_m^{(n)})$ converges to $\int_a^b f(x)\,dx$. For a system of nets which is not such that

$$\sum_{m=1}^{\infty} \delta_m^{(n)} F(\delta_m^{(n)})$$

converges to zero, $\sum_{m=1}^{m} \delta_m^{(n)} U(\delta_m^{(n)})$ does not in general converge, as $m \sim \infty$.

357. The definitions of $\int_a^\infty f(x)\,dx$, $\int_{-\infty}^b f(x)\,dx$ may be extended to the case in which $f(x)$ has points of infinite discontinuity. If the improper integral $\int_a^X f(x)\,dx$ exist for every value of X which is $> a$, and if it converge to a definite limit, as X increases indefinitely, then that limit defines

$$\int_a^\infty f(x)\,dx.$$

The integrals $\int_x^\infty f(x)\,dx$, $\int_{-\infty}^x f(x)\,dx$, when they exist, possess many of the properties of a proper, or an improper, integral $\int_a^x f(x)\,dx$. These integrals are continuous functions of the finite limit x. For

$$\int_{-\infty}^x f(x)\,dx = \int_{-\infty}^a f(x)\,dx + \int_a^x f(x)\,dx,$$

where $a < x$; and since $\int_a^x f(x)\,dx$ is a continuous function of x, so also is $\int_{-\infty}^x f(x)\,dx$.

When the integral $\int_a^\infty f(x)\,dx$ exists, the integral $\int_a^x f(x)\,dx$ is continuous for all values of x in the interval (a, ∞), including $x = \infty$, as it is there continuous in the extended sense of the term, when the point ∞ is regarded as belonging to the domain.

If the integral $\int_a^x f(x)\,dx$ exist for every finite value of x in the interval (a, ∞), and if $\phi(x)$ be a function which is finite and continuous for every such value of x, and be such that

$$\phi(x) - \phi(a) = \int_a^x f(x)\,dx,$$

then, provided $\phi(x)$ be continuous for $x = \infty$, the function $f(x)$ is integrable in (a, ∞), and $\phi(\infty) - \phi(a) = \int_a^\infty f(x)\,dx$. If the function $\phi(x)$ have a derivative, say $D^+\phi(x)$, which is integrable in every interval (a, x) of (a, ∞); then if the integral be a proper one, or be such an improper one that the relation

$$\phi(x) - \phi(a) = \int_a^x D^+\phi(x)\,dx$$

subsists, then, provided the limit $\phi(\infty)$ exist, we have also

$$\phi(\infty) - \phi(a) = \int_a^\infty D^+\phi(x)\,dx.$$

A similar statement applies to each of the other derivatives of $\phi(x)$. In the case in which $\phi(x) - \phi(a)$ differs from $\int_a^x D^+f(x)\,dx$ by an integrable null-function, this holds also for the limit $x = \infty$.

358. An integral $\int_a^\infty f(x)\,dx$ is said to be *absolutely convergent* when the integral $\int_a^\infty |f(x)|\,dx$ exists; otherwise it is said to be *conditionally* or *non-absolutely convergent*. If $\int_a^\infty |f(x)|\,dx$ exists, then also $\int_a^\infty f(x)\,dx$ exists; for $\left|\int_{x_1}^{x_2} f(x)\,dx\right| \leq \int_{x_1}^{x_2} |f(x)|\,dx$, and hence the convergence of the latter integral follows from that of the former one.

If $f(x)$ and $f(x)\,\phi(x)$ be both integrable in every interval (a, x), contained in (a, ∞), and if $\int_a^\infty f(x)\,dx$ be absolutely convergent, and if $\phi(x)$ be, from and after some fixed value of x, numerically less than some fixed number, then the integral $\int_a^\infty f(x)\,\phi(x)\,dx$ exists, and is absolutely convergent.*

---

* Riemann's *Ges. Werke*, p. 229; also Pringsheim, *Math. Annalen*, vol. XXXVII (1890), p. 591.

For, since $\int_a^\infty f(x)\,dx$ is absolutely convergent, we can find, corresponding to a fixed positive number σ, a number $\xi > a$, such that

$$\int_\xi^{\xi+h} |f(x)|\,dx < \sigma,$$

for all positive values of h; we have then

$$\left|\int_\xi^{\xi+h} f(x)\,\phi(x)\,dx\right| \leqq K\int_\xi^{\xi+h} |f(x)|\,dx \leqq K\sigma,$$

where K is the upper boundary of $|\phi(x)|$, and is by hypothesis finite. It is thus seen that $\int_a^\infty f(x)\,\phi(x)$ is convergent. Also since

$$\int_\xi^{\xi+h} |f(x)\,\phi(x)|\,dx \leqq K\int_\xi^{\xi+h} |f(x)|\,dx \leqq K\sigma,$$

we see that the convergence is absolute.

359. An important set of tests of the absolute convergence of an integral $\int_a^\infty f(x)\,dx$ is the following:

If $f(x)$ be integrable in every interval (a, x), then $\int_a^\infty f(x)\,dx$ converges to a definite finite value, provided $f(x)$ converge to zero, as x is increased indefinitely, in such a manner that one of the expressions

$$f(x).x^{1+k},\ f(x)\,x\,(\log x)^{1+k},$$
$$f(x)\,x\log x\,(\log\log x)^{1+k},\ \ldots\ldots f(x)\,x\log x.\log\log x\ \ldots\ldots\ (\log\log\ldots\log x)^{1+k},$$

converges to zero, as x is indefinitely increased, k denoting some fixed number greater than zero.

The integral $\int_a^\infty f(x)\,dx$ is not convergent, in case $f(x)$ be of invariable sign, from and after some fixed value of x, and provided also any one of the above expressions remains numerically greater than some fixed number (>0), as x is increased indefinitely, when k has the value zero.

We see that, in the first case, $\int_X^{X+h} |f(x)|\,dx$ is numerically less than one of the expressions

$$C\int_X^{X+h}\frac{dx}{x^{1+k}},\quad C\int_X^{X+h}\frac{dx}{x(\log x)^{1+k}},\quad C\int_X^{X+h}\frac{dx}{x\log x\,(\log\log x)^{1+k}},\ \ldots,$$

where C is a constant, dependent on X, which converges to zero as X is indefinitely increased. These expressions have the values

$$\frac{C}{k}\left[\frac{1}{X^k}-\frac{1}{(X+h)^k}\right],\quad \frac{C}{k}\left[\frac{1}{(\log X)^k}-\frac{1}{(\log X+h)^k}\right],$$
$$\frac{C}{k}\left[\frac{1}{(\log\log X)^k}-\frac{1}{(\log\log X+h)^k}\right],\ \ldots;$$

hence, k being positive, it is clear that X may be so chosen that

$$\int_X^{X+h} |f(x)|\, dx$$

is less than an arbitrarily fixed number, and thus

$$\int_a^{\infty} |f(x)|\, dx$$

is convergent. In the second case, k being now zero, we see that

$$\int_X^{X+h} f(x)\, dx$$

is numerically greater than one of the expressions

$$C\int_X^{X+h}\frac{dx}{x},\quad C\int_X^{X+h}\frac{dx}{x\log x},\quad C\int_X^{X+h}\frac{dx}{x\log x\log\log x},\ \dots$$

or than one of

$$C\log\frac{X+h}{X},\quad C\log\frac{\log(X+h)}{\log X},\quad C\log\log\frac{\log(X+h)}{\log X},\ \dots,$$

and these expressions increase indefinitely, as h is increased. It follows that $\int_a^{\infty} f(x)\, dx$ is in this case divergent.

CHANGE OF THE VARIABLE IN A SINGLE INTEGRAL

360. Let $f(x)$ be a function that is bounded in the interval (a, b). We now assume that x is a continuous monotone function, $\psi(\xi)$, of another variable ξ, defined for an interval (α, β), of ξ, and such that $a = \psi(\alpha)$, $b = \psi(\beta)$. The following theorem will be established:

If $D\psi(\xi)$, one of the derivatives of the non-diminishing monotone function $\psi(\xi)$, be integrable (R) in the interval (α, β), then

$$\overline{\int_a^b} f(x)\, dx = \overline{\int_\alpha^\beta} f\{\psi(\xi)\}\, D\psi(\xi)\, d\xi,\ \textit{and}\ \underline{\int_a^b} f(x)\, dx = \underline{\int_\alpha^\beta} f\{\psi(\xi)\}\, D\psi(\xi)\, d\xi.$$

Let a system of nets $\{\bar{D}_n\}$ be fitted on to the interval (α, β). Since the incrementary ratio $\dfrac{\psi(\xi_1) - \psi(\xi_2)}{\xi_1 - \xi_2}$ has, in any interval, the same upper and lower boundaries as $D\psi(\xi)$ in the same interval (see § 280), to a mesh $\bar{\delta}^{(n)}$ of $\bar{D}_n$ there corresponds a mesh $\delta^{(n)}$, of a net D_n, fitted on to (a, b), such that $\delta^{(n)} \leqq k\bar{\delta}^{(n)}$, where k is the upper boundary of $D\psi(\xi)$ in (α, β). If $\bar{u}^{(n)}$ denote the upper boundary of $f\{\psi(\xi)\}\, D\psi(\xi)$ in the interval $\bar{\delta}^{(n)}$, we see that $\bar{u}^{(n)}$ lies in the interval bounded by $U(\delta^{(n)})\, u^{(n)}$, and $U(\delta^{(n)})\, l^{(n)}$, where $u^{(n)}$, $l^{(n)}$ are the upper, and the lower, boundary of $D\psi(\xi)$ in the mesh $\bar{\delta}^{(n)}$. Also $\delta^{(n)}$ lies in the interval bounded by $u^{(n)}\,\bar{\delta}^{(n)}$ and $l^{(n)}\,\bar{\delta}^{(n)}$; we may therefore write

$$\delta_m^{(n)} = \bar{\delta}_m^{(n)}\{u_m^{(n)} - \theta_m^{(n)}(u_m^{(n)} - l_m^{(n)})\},\ \text{where } 0 \leqq \theta_m^{(n)} \leqq 1.$$

We have then $\Sigma\delta^{(n)}U(\delta^{(n)}) - \Sigma\bar{\delta}^{(n)}\bar{u}^{(n)}$

$$= \Sigma\bar{\delta}^{(n)}[\{u^{(n)} - \theta^{(n)}(u^{(n)} - l^{(n)})\}U(\delta^{(n)}) - \bar{u}^{(n)}]$$

$$= \Sigma\bar{\delta}^{(n)}\{u^{(n)}U(\delta^{(n)}) - \bar{u}^{(n)}\} - \Sigma\theta^{(n)}\bar{\delta}^{(n)}(u^{(n)} - l^{(n)})U(\delta^{(n)});$$

the summations being taken for the m_n meshes of the net $\bar{D}_n$, or of the net D_n. We shall assume that $f(x) \geqq 0$ in (a, b): by the theorem of § 346 this involves no loss of generality, for the constant C can be so chosen that $f(x) + C \geqq 0$. We then have

$$0 \leqq \Sigma\bar{\delta}^{(n)}\{u^{(n)}U(\delta^{(n)}) - \bar{u}^{(n)}\} \leqq \Sigma\bar{\delta}^{(n)}U(\delta^{(n)})(u^{(n)} - l^{(n)}) \leqq U\Sigma\bar{\delta}^{(n)}(u^{(n)} - l^{(n)}),$$

where U is the upper boundary of $f(x)$ in (a, b); we have also

$$0 \leqq \Sigma\theta^{(n)}\bar{\delta}^{(n)}(u^{(n)} - l^{(n)})U(\delta^{(n)}) \leqq \Sigma\bar{\delta}^{(n)}U(\delta^{(n)})(u^{(n)} - l^{(n)}) \leqq U\Sigma\bar{\delta}^{(n)}(u^{(n)} - l^{(n)}).$$

Since $D\psi(\xi)$ is integrable (R) in (α, β), it follows that $\Sigma\bar{\delta}^{(n)}(u^{(n)} - l^{(n)})$ is arbitrarily small, when n is sufficiently increased: it thus appears that

$$|\Sigma\delta^{(n)}U(\delta^{(n)}) - \Sigma\bar{\delta}^{(n)}\bar{u}^{(n)}|$$

is arbitrarily small. Therefore as $n \sim \infty$, the two sums

$$\sum_{m=1}^{m=m_n}\delta_m^{(n)}U(\delta_m^{(n)}), \quad \sum_{m=1}^{m_n}\bar{\delta}_m^{(n)}\bar{u}_m^{(n)}$$

converge to the same limit. Thus the theorem is proved for the case of the upper integrals, for the lower integrals it can be proved in a similar manner.

We deduce at once the following theorem for the transformation of a single integral:

If $D\psi(\xi)$, one of the derivatives of the monotone function $\psi(\xi)$, be integrable (R), in the interval (α, β), of ξ, and if either of the R-integrals

$$\int_a^b f(x)\,dx, \quad \int_\alpha^\beta f\{\psi(\xi)\}D\psi(\xi)\,d\xi$$

exists, the other also exists, and they have the same value; $f(x)$ being a bounded function.

Since $\psi(\xi)$ has a finite differential coefficient $\psi'(\xi)$, almost everywhere in the interval (α, β), and $\psi'(\xi)$ is integrable (R) in the interval (see § 345), provided one of the derivatives $D\psi(\xi)$ is integrable (R), we have

$$\int_a^b f(x)\,dx = \int_a^b f\{\psi(\xi)\}\psi'(\xi)\,d\xi,$$

provided either of the integrals is known to exist as an R-integral.

In case $\psi(\xi)$, although not monotone, is such that it is monotone in each of a finite number of intervals $(\alpha, \gamma_1), (\gamma_1, \gamma_2), \ldots (\gamma_n, \beta)$, where $\gamma_1, \gamma_2, \ldots$ do not necessarily all lie between α and β, these intervals of ξ may be considered separately, and the above result can be consequently extended to such a case.

The following theorem* for the transformation of a single R-integral s a particular case of the above theorem:

If the function $f(x)$ is integrable (R) in (a, b), and if a monotone function $\psi(\xi)$ *is defined by* $x = \psi(\xi) = c + \int_\alpha^\xi \phi(\xi)\,d\xi$, *where* $\psi(\xi)$ *varies from* α *to* β *as* x *varies from* a *to* b, *and* $\phi(\xi)$ *has an* R*-integral in* (α, β), *then*

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f\{\psi(\xi)\}\,\phi(\xi)\,d\xi.$$

361. It may happen that, when $\psi(\xi)$ is a monotone function of ξ, the nterval (a, b), of x, corresponds to the infinite interval (α, ∞), of ξ. We now assume that $\psi(\infty)$ is the limit of $\psi(\xi)$, as $\xi \sim \infty$, *i.e.* that $\psi(\xi)$ is continuous at ∞. If the conditions of the first theorem above are satisfied or every interval $(a, b-\epsilon)$, of x, with the corresponding interval (α, β'), of ξ, we have

$$\int_a^{b-\epsilon} f(x)\,dx = \int_\alpha^{\beta'} f\{\psi(\xi)\}\,D\psi(\xi)\,d\xi.$$

Since this holds for every value of ϵ, we have, on proceeding to the limit $= 0$,

$$\int_a^b f(x)\,dx = \int_\alpha^\infty f\{\psi(\xi)\}\,D\psi(\xi)\,d\xi,$$

n accordance with the definition in § 355, of the integral on the right-hand side. Similar considerations apply to the case in which the value of x corresponds to $\xi = -\infty$.

If it be desired to transform the integral $\int_a^b f(x)\,dx$, by means of the relation $\xi = \phi(x)$, where $\phi(x)$ is a single-valued function of x, then, unless $\phi(x)$ be monotone in the interval (a, b), the inverse function $\psi(\xi)$ will not be everywhere single-valued.

If it be assumed that $\phi(x)$ is monotone in (a, b), and that $\alpha = \phi(a)$, $\beta = \phi(b)$, a derivative $D\phi(x)$, of $\phi(x)$, is reciprocal to a derivative $D\psi(\xi)$ of $\psi(\xi)$. If it be assumed that $\frac{1}{D\phi(x)}$ is integrable in (α, β), and that the same holds for $\frac{f(x)}{D\phi(x)}$ considered as a function of ξ, or else that $f(x)$ is ntegrable in (a, b), then we may use the transformation

$$\int_a^b f(x)\,dx = \int_\alpha^\beta \left\{\frac{f(x)}{D\phi(x)}\right\}_{\xi=\phi(x)} d\xi.$$

If $\phi(x)$ be not monotone, $\int_a^b f(x)\,dx$ cannot in general be transformed nto a single integral in y. If, for example, $\phi(x)$ increases from $x = a$ to $x = k$, and then diminishes from $x = k$ to $x = b$, we must take

$$\int_a^b f(x)\,dx = \int_\alpha^{\phi(k)} \left\{\frac{f(x)}{D\phi(x)}\right\}_{\xi=\phi(x)} d\xi + \int_{\phi(k)}^\beta \left\{\frac{f(x)}{D\phi(x)}\right\}_{\xi=\phi(x)} d\xi,$$

---

* Lebesgue, *Annales de Toulouse*, (3), vol. I (1909), p. 44.

and the integrals on the right-hand side cannot in general be combined into one integral through the interval (α, β), because in the two integrals the integrand has different values for the same value of y.

Thus, for example, if $y = \sin x$,

$$\int_0^\pi f(x)\,dx = \int_0^1 \left[\frac{f(x)}{\cos x}\right] dy + \int_1^0 \left[\frac{f(x)}{\cos x}\right] dy$$

$$= \int_0^1 \frac{f(\sin^{-1} y)}{\sqrt{1-y^2}}\,dy + \int_0^1 \frac{f\left(\frac{\pi}{2} + \sin^{-1} y\right)}{\sqrt{1-y^2}}\,dy,$$

the value of $\cos x$ in the second integral on the right-hand side of the first equation being negative, and the value of $\sin^{-1} y$ being in the interval $(0, \frac{1}{2}\pi)$.

REPEATED INTEGRALS

362. The actual evaluation of a double integral over the fundamental rectangle, of which the sides are $x = x_0$, $x = x_1$, $y = y_0$, $y = y_1$, is usually made to depend upon the evaluation of successive single integrals taken, first with respect to one of the variables, and then with respect to the other. The expression

$$\int_{x_0}^{x_1} dx \int_{y_0}^{y_1} f(x, y)\,dy, \text{ or } \int_{x_0}^{x_1}\int_{y_0}^{y_1} f(x, y)\,dy\,dx,$$

in which $f(x, y)$ is supposed to be integrated first with respect to y, for a constant value of x, and then with respect to x, is called a repeated integral. Similarly, the expression

$$\int_{y_0}^{y_1} dy \int_{x_0}^{x_1} f(x, y)\,dx,$$

in which the integrations are performed in the reverse order, is called a repeated integral. The question of the existence of these repeated integrals, and, in any given case, their relation with one another, and with the double integral, will be here investigated. It will be observed that the double integral has been defined as a single limit; whereas the repeated integrals, when they exist, are each obtained as the results of repeated limits. We have then to investigate whether, or under what conditions, a double integral is capable of representation as a repeated limit, of one of the forms indicated. It cannot be assumed *a priori* that the existence of the double integral necessarily implies the existence, for each value of x, of the single integral

$$\int_{y_0}^{y_1} f(x, y)\,dy$$

as a definite number. Neither is the existence of this single.integral, as a definite number, necessary for the existence, as a definite number, of the repeated integral

$$\int_{x_0}^{x_1} dx \int_{y_0}^{y_1} f(x, y)\, dy.$$

In fact, if we assume that the upper and lower integrals

$$\overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

have different values for some of the values of x, it may happen that the two repeated limits

$$\int_{x_0}^{x_1} dx \overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \int_{x_0}^{x_1} dx \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

have identical values.

The repeated integral will consequently be regarded, in this case, as existing; and thus it may be defined as

$$\int_{x_0}^{x_1} dx \overline{\underline{\int_{y_0}^{y_1}}} f(x, y)\, dy,$$

where the upper or lower integral with respect to y is to be taken indifferently, provided the repeated limit exists as a definite number.

In a similar manner $\displaystyle\int_{y_0}^{y_1} dy \overline{\underline{\int_{x_0}^{x_1}}} f(x, y)\, dx,$

when it has a definite value independent of whether the upper or the lower integral with respect to x be used, will be regarded as the repeated integral, first with respect to x and then with respect to y.

363. It was first established by P. Du Bois-Reymond* that, when the bounded function $f(x, y)$ has a double R-integral in the fundamental rectangle, then the two repeated integrals exist, and are each equal to the double integral. We shall first give a proof† of this theorem which exhibits its relation with the theory of sets of points.

The following preliminary theorem will first be established:

If $\int f(x, y)\, d(x, y)$, *taken over the fundamental rectangle, have a definite value, then the values of* x *for which the single integral* $\int f(x, y)\, dy$, *taken with* x *constant, has a definite value, define a set of points on a side of the rectangle, of linear measure equal to the length of that side.*

---

* *Crelle's Journal*, vol. xciv (1883), p. 277.

† The investigation is founded on that of Schoenflies, *Bericht*, vol. I, p. 193.

It follows from this theorem, that the set is everywhere dense in the interval, and of cardinal number c. Moreover, the points at which

$$\overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

differ from one another form a set of measure zero.

On the assumption of the existence of the double integral, the set K of all the points at which the saltus of $f(x, y)$ is $\geq k$, where k is an arbitrarily chosen positive number, is a closed set, of plane content zero. If a straight line be drawn parallel to the y-axis through the point x of the side $y = y_0$, of the rectangle, then the component of K on this straight line will be denoted by K_x, and its linear content by $I(K_x)$. It has been shewn in § 143 that, σ denoting a prescribed positive number, the linear content of that set of points x, on the side $y = y_0$, of the rectangle, for which $I(K_x) \geq \sigma$, is zero; and thus that $I(K_x)$, considered as a function of x, is an integrable null-function, for each value of k. The function $\chi(x), = \lim\limits_{k \sim 0} I(K_x)$, is also an integrable null-function: for the set of points at which $\chi(x)$ does not vanish is made up of those sets of points at which $I(K_x^{(1)}), I(K_x^{(2)}), \ldots I(K_x^{(n)}), \ldots$ do not vanish: where

$$K^{(1)}, K^{(2)}, \ldots K^{(n)}, \ldots$$

correspond to a diminishing sequence of values of k converging to the limit zero; and since each of these sets has zero measure, it follows that the set of points at which $\chi(x)$ does not vanish has zero measure. At any point x_1, at which $\chi(x)$ vanishes, $I(K_{x_1})$ vanishes, for every value of k.

It should be observed that, at a point of K_x, it is not necessarily the case that $f(x, y)$, considered as a function of y, with x constant, has its saltus $\geq k$; in fact, this saltus may be less than k, or may be zero. However, all the points at which the saltus of $f(x, y)$, taken with x constant, is $\geq k$, are certainly included in the set K_x.

For any fixed value of x, the upper and lower integrals

$$\overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

both exist, and the two have equal values at any point of the everywhere dense set of points x, of linear measure $x_1 - x_0$, at which $\chi(x)$ vanishes. The preliminary theorem above stated has thus been established.

Let $F(x)$ denote $\int_{y_0}^{y_1} f(x, y)\, dy$, where $F(x)$ consequently has a single determinate value at each point x, at which

$$\overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

are equal. At any point of that set, of measure zero, at which

$$\overline{\int_{y_0}^{y_1}} f(x, y)\, dy, \quad \underline{\int_{y_0}^{y_1}} f(x, y)\, dy$$

have different values, $F(x)$ is regarded as indeterminate; and the upper and lower integrals are the upper and lower limits of indeterminancy. It will now be shewn that the function $F(x)$, so defined, is an integrable function.

Let x' be a point on the side $y = y_0$, of the rectangle $(x_0, y_0; x_1, y_1)$, such that the component $K_{x'}$, of K, on the line $x = x'$, has content $< \sigma$. A finite number of intervals $\epsilon_1, \epsilon_2, \ldots \epsilon_m$ can be determined on the straight line $x = x'$, neither abutting on, nor overlapping, one another, such that their sum $\epsilon_1 + \epsilon_2 + \ldots + \epsilon_m > b - \sigma$, where b is the length of the side of the rectangle parallel to the y-axis, and such also as to contain, in their interiors and at their ends, no points at which the saltus of $f(x, y)$ is $\geqq k$. For each point of one of these intervals ϵ there exists a rectangle, with the point at the centre, such that the fluctuation of $f(x, y)$ in that rectangle is $< k$. The breadths of these rectangles for all points of ϵ must have a finite minimum, for otherwise there would exist a point of ϵ which would belong to K. It follows that, for the point x', an interval $(x' - \alpha, x' + \beta)$ can be determined, such that the straight lines $x = x' - \alpha$, $x = x' + \beta$ intersect all the rectangles corresponding to all the points of the intervals $\epsilon_1, \epsilon_2, \ldots \epsilon_m$. If x_1, x_2 be any two points in the interval $(x' - \alpha, x' + \beta)$, we have

$$| F(x_1) - F(x_2) | < bk + \sigma (U - L).$$

Now a finite number of separate intervals $\delta_1, \delta_2, \ldots \delta_r$ can be determined on the side $y = y_0$, of the rectangle (length $= a$), such that

$$\delta_1 + \delta_2 + \ldots + \delta_r > a - \eta,$$

where η is a prescribed positive number, and such that each point of each of the intervals δ is a point x', for which an interval $(x' - \alpha, x' + \beta)$ can be determined as above. By applying the Heine-Borel theorem we see that

$$\delta_1, \delta_2, \ldots \delta_r$$

will all be covered by a finite number of these intervals $(x' - \alpha, x' + \beta)$. It thus appears that the side $y = y_0$, of the rectangle, can be divided into a finite number of parts

$$\tau_1, \tau_2, \ldots \tau_p,$$

and

$$\lambda_1, \lambda_2, \ldots \lambda_q,$$

such that

$$\tau_1 + \tau_2 + \ldots + \tau_p > a - \eta,$$

and

$$\lambda_1 + \lambda_2 + \ldots + \lambda_q < \eta,$$

so that the fluctuation of $F(x)$ in any one of the parts τ is

$$< bk + \sigma (U - L).$$

Let $$k = \epsilon/2b,\ \sigma = \epsilon/2\,(U - L).$$
then we see that $F(x)$ is such that the side $y = y_0$, of the fundamental rectangle, can be divided into a finite number of parts, such that the sum of those parts, in which the fluctuation of $F(x)$ is $\geq \epsilon$, is less than the arbitrarily chosen number η. It follows that $F(x)$ is integrable along the side of the rectangle.

It has now been shewn, on the assumption of the existence of the double integral, that the repeated integral

$$\int_{x_0}^{x_1} F(x)\,dx, \text{ or } \int_{x_0}^{x_1} dx \overline{\underline{\int}}_{y_0}^{y_1} f(x, y)\,dy,$$

taken over the fundamental rectangle, has a definite value. Moreover, this value is equal to that of the double integral. For, let the fundamental rectangle be divided up by means of straight lines parallel to the y-axis, through the end-points of each interval of the two finite sets $\{\tau\}$ and $\{\lambda\}$. Any one of the rectangles so constructed, with τ as base and with height b, can be divided into parts, by means of straight lines parallel to the x-axis, such that, in each one of a number of these parts the sum of whose heights is $> b - \sigma$, the fluctuation of $f(x, y)$ is $< k$. The fundamental rectangle has now been divided into a finite number of parts, such that the sum of the products of each part, multiplied by the upper boundary of $f(x, y)$ in that part, exceeds the sum of the products of each part, multiplied by the lower boundary of the function in that part, by less than

$$abk + (a\sigma + b\eta)(U - L),$$

which is arbitrarily small.

Also $$\int_{x_0}^{x_1} dx \overline{\underline{\int}}_{y_0}^{y_1} f(x, y)\,dy, \text{ and } \int f(x, y)\,d(x, y),$$

both lie between the two sums of products, and therefore differ from one another by less than $abk + (a\sigma + b\eta)(U - L)$. The equality of the double integral and the repeated integral is thus put in evidence by the mode of sub-division of the rectangle which has been adopted.

Similar reasoning applies to the repeated integral in which the integration is taken first with respect to x, and then with respect to y.

It has thus been established that, *if the double integral through the fundamental rectangle exist, then the two repeated integrals also exist, and are each equal to the double integral.*

All the points at which $\chi(x)$ vanishes are points of continuity of the function $F(x)$; but there may also be other points at which $F(x)$ is continuous; because the existence of a saltus of $f(x, y)$ at a point (x, y) is consistent with $f(x, y)$ being continuous with respect to x, and also with respect to y, at the point.

The function $F(x)$ may be replaced by $\psi(x)$, the most nearly continuous function related to it (see § 242). We thus have

$$\int f(x, y)\, d(x, y) = \int_{x_0}^{x_1} \psi(x)\, dx.$$

If it be assumed that the set of points G, for which $f(x, y)$ is defined, is measurable in accordance with Jordan's definition of a measurable set (see § 142), then the double integral of the bounded function may be replaced by that of the function $f(x, y)$, defined for all points of a rectangle which contains G, by the convention that $f(x, y)$ shall vanish at all those points of the rectangle which do not belong to G. If the set G be such that each straight line parallel to the y-axis contains points of G which fill up a finite, or an indefinitely great number of continuous intervals, or more generally, if the set of such points, for each value of x, be linearly measurable (J), then the integral

$$\int f(x, y)\, dy,$$

taken along the whole segment of the line between the sides of the rectangle, may be replaced by the same integral taken through the component of G on the same segment. In particular, if the points of G on the straight line through the point x consist of all the points in the linear interval

$$(f_1(x), f_2(x)),$$

we may replace

$$\int_{y_0}^{y_1} f(x, y)\, dy$$

by

$$\int_{f_1(x)}^{f_2(x)} f(x, y)\, dy;$$

and therefore in this case,

$$\int_G f(x, y)\, d(x, y) = \int_{x_0}^{x_1} dx \int_{f_1(x)}^{f_2(x)} f(x, y)\, dy.$$

364. A simple proof* of the fundamental theorem of § 363 will be given, which depends upon the fact that, for any bounded function $f(x, y)$, if the operation of taking the upper integral first with respect to y, and then with respect to x, be performed, the result cannot exceed the upper double integral; and that, similarly, the result of successively taking the lower integral with respect to x and to y cannot be less than the lower double integral: thus

---

* This method of proof was first employed by Harnack; see his edition of Serret's *Differential and Integral Calculus*, p. 282; also *Math. Annalen*, vol. XXVI (1886), p. 566. Other proofs of this kind have been given by Arzelà, *Mem. dell' Ist. di Bologna* (5), vol. II (1891), p. 133; by Jordan, *Liouville's Journal* (4), vol. VIII (1892), p. 84, or *Cours d'Analyse*, vol. I, p. 42; also by Pringsheim, *Sitzungsberichte d. Münch. Akad.* vol. XXVIII (1898), p. 59, and vol. XXIX (1899), p. 39; the first of these contains a history of the literature of the theorem. See also Pierpont's paper "On multiple integrals," *Trans. Amer. Math. Soc.* vol. VI (1905), where a proof of this character for multiple integrals is given.

$$\underline{\int} f(x, y)\, d(x, y) \leq \underline{\int} dx \underline{\int} f(x, y)\, dy \leq \overline{\int} dx \overline{\int} f(x, y)\, dy$$

$$\leq \overline{\int} f(x, y)\, d(x, y),$$

the integrals being all taken over the fundamental rectangle.

If $f(x, y)$ be integrable, so that

$$\overline{\int} f(x, y)\, d(x, y) = \underline{\int} f(x, y)\, d(x, y),$$

it follows that $$\overline{\int} dx \overline{\int} f(x, y)\, dy = \overline{\int} dx \underline{\int} f(x, y)\, dy$$

$$= \underline{\int} dx \overline{\int} f(x, y)\, dy = \underline{\int} dx \underline{\int} f(x, y)\, dy,$$

and thus that the repeated integral

$$\int dx \int f(x, y)\, dy$$

has a definite value equal to the double integral.

To prove this, let the rectangle be divided into a number of parts δ, by means of straight lines parallel to the sides. Since the double integral is assumed to exist, this may be done in such a manner that, ϵ denoting an arbitrarily chosen positive number, the conditions

$$\int f(x, y)\, d(x, y) - \epsilon \leqq \Sigma\{\delta L(\delta)\} \leqq \Sigma\{\delta U(\delta)\} \leqq \int f(x, y)\, d(x, y) + \epsilon$$

are satisfied, where the summation Σ is taken for all the rectangles δ, and $U(\delta)$, $L(\delta)$ denote the upper and lower boundaries of $f(x, y)$ in a rectangle δ. Now, if we take the upper and lower integrals of $f(x, y)$ along a straight line parallel to the y-axis, we have

$$\overline{\int} f(x, y)\, dy \leqq \Sigma_1\{\delta' U(\delta)\},$$

$$\underline{\int} f(x, y)\, dy \geqq \Sigma_1\{\delta' L(\delta)\},$$

where the summation Σ_1 refers to all those rectangles δ which are intersected by the straight line along which the upper and lower integrals are taken; and in the case when that straight line is along one or more boundaries of the rectangles δ, Σ refers to all the rectangles on one side of that line: also δ' denotes that interval along the line of integration which is in the rectangle δ. It follows that

$$\overline{\int} dx \overline{\int} f(x, y)\, dy \leqq \Sigma \{\delta U(\delta)\} \leqq \overline{\int} f(x, y)\, d(x, y) + \epsilon$$

and
$$\underline{\int} dx \underline{\int} f(x, y)\, dy \geqq \Sigma \{\delta L(\delta)\} \geqq \underline{\int} f(x, y)\, d(x, y) - \epsilon;$$

and since these inequalities hold for every value of ϵ, we have

$$\overline{\int} dx \overline{\int} f(x, y)\, dy \leqq \overline{\int} f(x, y)\, d(x, y),$$

$$\underline{\int} dx \underline{\int} f(x, y)\, dy \geqq \underline{\int} f(x, y)\, d(x, y);$$

and thus the theorem is established.

365. The converse questions now arise whether, from the existence of one of the repeated integrals, or from the existence and equality of both repeated integrals, that of the double integral can be inferred. The answer to both questions must be in the negative. Continuity of a function $f(x, y)$ with respect to x and y separately does not necessarily imply continuity with respect to (x, y); moreover, the saltus of the function at a point with respect to x, when y has a constant value, or with respect to y, when x has a constant value, is not necessarily equal to the saltus of the function with respect to (x, y). It may happen that the component of K on a straight line parallel to one of the axes may consist of points, some or all of which are points of continuity of the function when considered as a function of one variable on that straight line. Thus K may have a plane content greater than zero; and yet the linear content of the points on all straight lines parallel to the axes, at which the linear saltus of the function is $\geqq k$, may be zero. Hence either, or both, of the repeated integrals may exist, whilst for values of k, the sets K are not of content zero*; and therefore whilst $f(x, y)$ does not admit of a double R-integral.

It was shewn† by W. H. Young that the existence of the R-integrals $\int_0^1 f(x, y)\, dx$, for all values of y such that $0 \leq y \leq 1$, and the existence of the R-integral $\int_0^1 f(x, y)\, dy$, for all values of x such that $0 \leq x \leq 1$, the function $f(x, y)$ being assumed to be bounded in square $(0, 0;\ 1, 1)$, necessitate the existence of the two repeated integrals. It was further shewn* by Lichtenstein that, subject to the same conditions, the two repeated integrals exist and have the same value. A further investigation is that† of D. C. Gillespie. The case of multiple integrals has been treated‡ by Ettlinger.

---

* An incorrect theorem, relating to this point, has been given by Schoenflies, see his *Bericht*, vol. I, p. 197. In this theorem, the condition that K should be closed is stated to be the condition for the existence of the double integral. If, however, K were not closed, it could not represent the set of points at which any function had a saltus $\geqq k$ The examples given by Schoenflies do not in reality accord with his theorem.

† *Monatshefte für Math u. Physik*, vol. XXI (1910), p. 127.

EXAMPLES

1§. For the rectangle bounded by $x = 0$, $x = 1$, $y = 0$, $y = 1$, let $f(x, y) = 1$, for all rational values of x, and $f(x, y) = 2y$, for all irrational values of x. We have then

$$\int_0^1 f(x, y)\, dy = 1,$$

whatever value x may have; and hence the repeated integral

$$\int_0^1 dx \int_0^1 f(x, y)\, dy$$

has the value 1: but the double integral does not exist, since $I(k) > 0$, for any value of k in the interval $(0, 1)$.

2‖. Let x be represented by a finite or infinite decimal, excluding those decimals in which every figure from and after some fixed place is 9. Let p_x denote the number of decimal places in the representation of x in the manner described. Let y be represented in a similar manner, with a corresponding definition of p_y. Let the function $f(x, y)$ be defined in the rectangle bounded by $x = 0$, $x = 1$, $y = 0$, $y = 1$, by $f(x, y) = \frac{1}{p_x + 1} + \frac{1}{p_y + 1}$ when p_x and p_y are both finite, otherwise let $f(x, y) = 0$.

We have $\int_0^1 \frac{1}{p_y + 1}\, dy = 0$; for there is only a finite number of values of y, in $(0, 1)$, for which p_y is less than an arbitrarily chosen fixed integer, or $\frac{1}{p_y + 1}$ is greater than an arbitrarily chosen fixed proper fraction. The function $f(x, y)$ vanishes, except when one at least of x and y is representable by a finite decimal; and thus the double integral

$$\int f(x, y)\, d(x, y) \text{ vanishes.}$$

Now $$\overline{\int_0^1} f(x, y)\, dy = \frac{1}{p_x + 1}, \quad \underline{\int_0^1} f(x, y)\, dy = 0;$$

and thus $\int_0^1 f(x, y)\, dy$ has no definite value for any value x of the everywhere dense enumerable set of points for which p_x is finite.

Nevertheless $$\int_0^1 dx \underline{\overline{\int_0^1}} f(x, y)\, dy = 0 = \int f(x, y)\, d(x, y).$$

3‖. With the same notation as in the last example, let $f(x, y) = 0$, when p_x, p_y are both finite, or both infinite; let $f(x, y) = \frac{1}{1 + p_x}$, when p_x is finite, and p_y infinite; and $f(x, y) = \frac{1}{1 + p_y}$, when p_y is finite, and p_x is infinite. In this case $f(x, y)$ differs from 0 at an unenumerable set of points; and yet the set of points at which $f(x, y) > \epsilon$ has the plane content zero, since all such points are on a finite number of lines parallel to the coordinate axis, although they are everywhere dense on those lines. The double integral, and consequently the repeated integrals, exist in this case.

---

* *Göttinger Nachrichten*, 1910, p 468; also *Sitzber. Berl. math. Ges.* vol. x (1911), p. 55.

† *Annals of Math.* (2), vol. xx (1919), p 224.

‡ *Amer. Journal of Math.* vol. XLVII (1926), p. 215.

§ Thomae, *Schlömilch's Zeitschrift*, vol. XXIII (1878), p. 67.

‖ Pringsheim, *Sitzungsber. d. Leipziger Akad* vol. XXVIII (1898), p. 71.

4[*]. An example has been given in Ex. 1. § 143, of a set of points K which is everywhere dense and unclosed, whereas the sets K_x, K_y are all finite. Let $f(x, y) = c'$, at every point of K, and $= c$, at every other point. In this case, the double integral does not exist; but

$$\int_0^1 f(x, y)\, dx = c, \quad \int_0^1 f(x, y)\, dy = c,$$

whatever values y and x may have in the first, and in the second, integral respectively.

Consequently $\int_0^1 dx \int_0^1 f(x, y)\, dy$ and $\int_0^1 dy \int_0^1 f(x, y)\, dx$

both exist, and have the same value c.

5[*]. Let a set $\{(x', y')\}$ be defined as follows:—Let x' have any value for which $p_{x'}$ is finite; and with such a fixed value of x', let every y' be taken for which $p_{y'} \leqq p_{x'}$. On every line parallel to the y-axis there is only a finite number of points of the set; but the set is everywhere dense on every line which is parallel to the x-axis, and which has for its ordinate one of the y'. Let $f(x, y) = c'$, for the set $\{(x', y')\}$, and let $f(x, y) = c$, for all remaining points.

We have, in this case,

$$\int_0^1 f(x, y)\, dy = c, \text{ and } \int_0^1 dx \int_0^1 f(x, y)\, dy = c.$$

But $\int_0^1 f(x, y')\, dx$ has c and c' for its upper and lower values; and the set of values of y' being everywhere dense, $\int_0^1 dy \int_0^1 f(x, y)\, dx$ does not exist.

6[†]. Let $f(x, y)$ be defined at all points of the rectangle bounded by $x = 0$, $x = 1$, $y = 0$, $y = 1$, by the condition that $f(x, y) = 0$, except at those points (x', y') at which

$$x' = \frac{2m + 1}{2^n}, \quad y' = \frac{2p + 1}{2^q}, \text{ where } f(x', y') = \frac{1}{2^n},$$

m, n, p and q being positive integers.

In this case the double integral exists, and therefore the repeated integrals both exist[‡].

IMPROPER DOUBLE INTEGRALS

366. As in the case of single integrals, the definition of a double integral may be extended to the case in which the function has a set of points of infinite discontinuity. This set is necessarily closed, and it will be assumed throughout that its plane content is zero. It will also be assumed that the domain for which such a function is defined is bounded, and that the frontier has the content zero, the domain being therefore measurable in accordance with Jordan's definition; and consequently the function may be replaced by another function defined for all the points in a fundamental rectangle, the new function being taken to vanish at all points not in the original domain, and to have the same value as the original function at all points of that domain. These assumptions being made, a definition of the improper double integral which is substantially the one given by Jordan*, and adopted by Stolz†, may be stated as follows:

---

[*] Pringsheim, *Sitzungsber. d. Leipziger Akad.* vol. XXVIII (1898), p. 71.

[†] Du Bois-Reymond, *Crelle's Journal*, vol. XCIV, p. 278; also Stolz, *Grundzüge*, vol. III (1899), p. 73.

[‡] This is denied by Stolz, *Grundzüge*, vol. III, p. 88, on the ground that $f(x, y)$ for $x = \frac{2m+1}{2^n}$ is not integrable with respect to y, but has $\frac{1}{2^n}$ and 0 for its upper and lower values. We have, however, shewn that this is no justification for denying the existence of the repeated integral.

Let $D_1, D_2, \ldots D_n, \ldots$ *denote a sequence of domains contained in the fundamental rectangle, each one of which is contained in the next and consists of a closed set with its frontier of content zero. Further, let us suppose that none of these domains contain, in their interiors or on their frontiers, any point at which* $f(x, y)$ *has an infinite discontinuity, and that the sequence is such that the measure of* D_n *as* $n \sim \infty$, *converges to that of the fundamental rectangle; then if the upper integrals*

$$\overline{\int}_{D_1} f(x, y)\, d(x, y),\quad \overline{\int}_{D_2} f(x, y)\, d(x, y), \ldots \overline{\int}_{D_n} f(x, y)\, d(x, y), \ldots,$$

taken over the domains $D_1, D_2, \ldots$, *converge to a definite limit, independent of the particular sequence* $\{D_n\}$ *chosen, this limit is defined to be the improper upper integral*

$$\overline{\int} f(x, y)\, d(x, y)$$

of $f(x, y)$ *in the given domain. A similar statement applies to the case of the improper lower integral. When the improper upper and lower integrals both exist, and have the same value, then the improper integral*

$$\int f(x, y)\, d(x, y)$$

over the given domain is said to exist, and to have this common value.

It will be observed that the domains D_n are all measurable in accordance with Jordan's definition of a measurable set, and therefore also in accordance with the definition of Borel and Lebesgue.

In case the function $f(x, y)$ be integrable (R) in all the domains $D_1, D_2, \ldots$, however this sequence may be chosen, subject to the conditions stated above, then, if the sequence

$$\int_{D_1} f(x, y)\, d(x, y),\quad \int_{D_2} f(x, y)\, d(x, y), \ldots$$

converge to a definite limit, independent of the particular sequence $\{D_n\}$, that limit defines the improper double integral

$$\int f(x, y)\, d(x, y).$$

---

* *Cours d'Analyse*, vol. II, 2nd ed. (1894), p. 76.

† *Grundzuge*, vol. III (1899), p. 124.

If a function $f(x, y)$ have an improper integral in the fundamental rectangle, then $f(x, y)$ has a proper R-integral in any closed domain of which the frontier has measure zero, and which is contained in the fundamental rectangle, but itself contains no points in its interior, or on its frontier, at which $f(x, y)$ is infinitely discontinuous. For, if D be such a domain, we may choose a sequence $\{D_n\}$, such as is employed in the definition given above, and such that D is interior to D_n, for every value of n.

Since $$\overline{\int}_D f(x, y)\, d(x, y) - \underline{\int}_D f(x, y)\, d(x, y)$$

cannot exceed $$\overline{\int}_{D_n} f(x, y)\, d(x, y) - \underline{\int}_{D_n} f(x, y)\, d(x, y)$$

whatever value n may have, and since this latter expression converges to zero, as $n \sim \infty$, it follows that

$$\overline{\int}_D f(x, y)\, d(x, y) - \underline{\int}_D f(x, y)\, d(x, y) = 0,$$

or that $\int_D f(x, y)\, d(x, y)$ exists.

It has thus been shewn that *if $f(x, y)$ have an improper double integral in the fundamental rectangle, it must possess a proper R-integral in any closed domain interior to that rectangle, such that the domain has its frontier of zero measure, and contains no points of infinite discontinuity of the function, either in its interior or on its frontier.*

In particular $f(x, y)$ has a proper R-integral over each of the domains D_n.

367. *The necessary and sufficient condition for the existence of the improper upper double integral*

$$\overline{\int} f(x, y)\, d(x, y)$$

is that, corresponding to any arbitrarily chosen positive number ϵ, another positive number δ can be determined, such that, if Δ be any closed domain whatever, of which the frontier has measure zero, and which is contained in the fundamental rectangle, but itself contains no point of infinite discontinuity of $f(x, y)$, either in its interior or on its frontier, then, provided the measure of Δ is $< \delta$, the condition

$$\left| \overline{\int} f(x, y)\, d(x, y) \right| < \epsilon,$$

taken over Δ, is satisfied.

A similar theorem applies to the improper lower double integral.

To shew that the condition stated in the theorem is sufficient, let D, D' be two domains of the kind specified in § 366, such that $m(D)$, $m(D')$ both differ from the area of the fundamental rectangle by less than δ: they are both interior to the fundamental rectangle, and contain none of the points of infinite discontinuity of the function. Let d be the set of points of D which do not belong to D', and d' the set of points of D' which do not belong to D; then

$$m(d) < A - m(D') < \delta,$$

and $$m(d') < A - m(D) < \delta,$$

where A is the area of the fundamental rectangle. Also, since the domains $D + d'$, $D' + d$ are identical, we have

$$I_D - I_{D'} = I_d - I_{d'},$$

where I denotes the upper double integral

$$\overline{\int} f(x, y)\, d(x, y)$$

taken over the domain indicated by a suffix. It follows that

$$| I_D - I_{D'} | \leqq | I_d | + | I_{d'} | < 2\epsilon;$$

hence it is easily seen that any two sequences

$$\{I_{D_n}\}, \quad \{I_{D'_n}\}$$

both converge to one and the same definite limit.

To shew that the condition stated in the theorem is necessary, let us suppose that it is not satisfied. We thus assume that a domain d, of arbitrarily small measure, can be found, such that $| I_d | > \epsilon$.

Let D be interior to the rectangle, and such that

$$A - m(D) < \delta.$$

Taking D to contain d, we then have

$$m(D - d) > A - 2\delta,$$

provided $$m(d) < \delta.$$

The two domains D, $D - d$ both converge to A, if δ be decreased indefinitely; and
$$I_D - I_{D-d} = I_d;$$

thus $$| I_D - I_{D-d} | > \epsilon,$$

however small δ may be; hence the limit does not exist in this case.

The necessary and sufficient condition that the improper upper and lower integrals of $f(x, y)$ in the fundamental rectangle may both exist is that the improper upper integral of $| f(x, y) |$ may exist.

To shew that the condition stated is sufficient, we observe that, on the assumption of the existence of

$$\overline{\int} \mid f(x, y) \mid d(x, y),$$

it follows from the theorem established above that the upper integral of

$$\mid f(x, y) \mid$$

through a domain D of the type defined in § 366, tends to the limit zero, as $m(D)$ does so. Also

$$\left| \overline{\int}_D f(x, y)\, d(x, y) \right|, \quad \left| \underline{\int}_D f(x, y)\, d(x, y) \right|$$

are both less than, or equal to

$$\overline{\int}_D \mid f(x, y) \mid d(x, y),$$

as is easily seen. It follows that both the integrals

$$\overline{\int}_D f(x, y)\, d(x, y), \quad \underline{\int}_D f(x, y)\, d(x, y)$$

converge to zero, as $m(D)$ does so, and uniformly for all such domains D; and these are the sufficient conditions for the existence of

$$\overline{\int} f(x, y)\, d(x, y), \quad \underline{\int} f(x, y)\, d(x, y).$$

To prove that the condition stated is a necessary one, let us assume that, for every domain D, satisfying the specified conditions, and such that $m(D) < \delta$, we have

$$\left| \overline{\int}_D f(x, y)\, d(x, y) \right| < \epsilon.$$

Now let $$f(x, y) = f^+(x, y) - f^-(x, y),$$
where $$f^+(x, y) = f(x, y)$$
at all points where $f(x, y)$ is positive, and everywhere else let

$$f^+(x, y) = 0;$$

also $$f^-(x, y) = -f(x, y),$$
at every point where $f(x, y)$ is negative, and everywhere else $f^-(x, y)$ is zero. The domain D may be divided into a finite number of rectangles δ, some of which may lie partly outside D; the functions being taken to be zero in all such outlying portions. Denoting by U_δ the upper boundary of a function in the rectangle δ, we have

$$U_\delta\{f(x, y)\} = U_\delta\{f^+(x, y)\},$$

in all elements δ in which $f(x, y)$ has positive values; and in all other elements

$$U_\delta\{f^+(x, y)\} = 0.$$

The elements may be taken such that, if η be an arbitrarily chosen number, the inequalities

$$\Sigma\delta U_\delta\{f(x, y)\} - \int_D f(x, y)\, d(x, y) < \eta,$$

$$\Sigma\delta U_\delta\{f^+(x, y)\} - \overline{\int}_D f^+(x, y)\, d(x, y) < \eta$$

are both satisfied. These conditions are also satisfied for any domain contained in D.

We have now

$$\begin{aligned}\int_D f^+(x, y)\, d(x, y) \leqq \sum_D \delta U_\delta\{f^+(x, y)\} &\leqq \sum_{D_1} \delta U_\delta\{f(x, y)\} \\ &\leqq \overline{\int}_{D_1} f(x, y)\, d(x, y) - \eta \\ &< \epsilon + \eta,\end{aligned}$$

where D_1 consists of the domain which is composed of those elements δ, of D, in which $f(x, y)$ has positive values. Since η is arbitrarily small, we have

$$\overline{\int}_D f^+(x, y)\, d(x, y) \leqq \epsilon.$$

Again, since

$$\underline{\int}_D f(x, y)\, d(x, y) = -\overline{\int}_D -\{f(x, y)\}\, d(x, y),$$

we see that
$$\overline{\int}_D -\{f(x, y)\}\, d(x. y)$$

has the limit zero, when $m(D)$ has the limit zero; and from this it follows, as before, that

$$\overline{\int}_D f^-(x, y)\, d(x, y) \leqq \epsilon.$$

Since
$$|f(x, y)| = f^+(x, y) + f^-(x, y),$$

we have
$$\overline{\int}_D |f(x, y)|\, d(x, y) \leqq 2\epsilon;$$

and therefore
$$\overline{\int}_D |f(x, y)|\, d(x, y)$$

has the limit zero, when $m(D)$ converges to zero. It now appears, by employing the first theorem of this section, that

$$\overline{\int} |f(x, y)|\, d(x, y),$$

taken throughout the fundamental rectangle, has a definite value.

It should be observed that since

$$\underline{\int}_D |f(x, y)|\, d(x, y) \leqq \overline{\int}_D |f(x, y)|\, d(x, y),$$

the lower integral $$\underline{\int}_D |f(x, y)|\, d(x, y)$$

has the limit zero, whenever zero is the limit of

$$\overline{\int}_D |f(x, y)|\, d(x, y);$$

and that therefore $$\underline{\int} |f(x, y)|\, d(x, y)$$

always exists when $$\overline{\int} |f(x, y)|\, d(x, y)$$

does so, the integration being over the fundamental rectangle.

368. It has been seen in § 366 that, in case $f(x, y)$ have an improper integral in the fundamental rectangle, it has a proper integral in any closed domain D contained in that rectangle, which has a frontier of zero measure, and contains no points of infinite discontinuity of the function, either interior to it, or on its frontier. It follows by the property (2) of § 340, that $|f(x, y)|$ is also integrable in the domain D; and we have already seen that the existence of the improper upper integral of $|f(x, y)|$ is a necessary consequence of the existence of the improper integral of $f(x, y)$. It thus appears that the improper upper and lower integrals of $|f(x, y)|$ must be identical, and therefore that, *if $f(x, y)$ be a function which has an improper double integral, in accordance with Jordan's definition, then $|f(x, y)|$ has also an improper integral, so that every such improper integral is absolutely convergent.*

We have seen that Cauchy's definition of an improper single integral is applicable not only to the cases in which the convergence is absolute, but also to cases in which the convergence is not absolute. The same remark applies to Harnack's extension of Cauchy's definition, which will be considered in Chapter VII. Jordan's definition of an improper double integral is however much more stringent than Harnack's definition of an improper single integral. In the latter case the integral is defined as the limit of the proper integral taken through a finite number of intervals, not chosen arbitrarily in any manner consistent with the condition that the sum of these intervals is to converge to the length of the interval of integration, but chosen so as to satisfy the special condition that they are complementary to a finite set of intervals which contain within them all the points of infinite discontinuity of the function, each interval of the finite set containing at least one such point.

If the proper integrals, of which the improper integral, in Harnack's definition, is the limit, were not subjected to the above-mentioned restriction, reasoning precisely similar to that applied above would shew that every improper single integral must be absolutely convergent. In order that a definition of the improper double integral should admit of the existence of double integrals which do not converge absolutely, it would be necessary to subject the domains $D_1, D_2, \dots D_n, \dots$ (the proper integrals through which define, as their limit, the double integral), to some restriction which would allow of the existence of a limit in cases in which such a limit does not otherwise exist, independently of the particular set $\{D_n\}$ chosen. Such a restriction as to the nature of the domains D_n would correspond to the restriction to a special class of sets of intervals, of the intervals through which the proper integrals in Cauchy's or Harnack's definition of an improper single integral are taken. The true extension of Harnack's definition to the case of double integrals would be the following:

Let the points of infinite discontinuity of $f(x, y)$ (the set of such points being of zero content) be enclosed in a finite set of rectangles with sides parallel to those of the fundamental rectangle, each rectangle of the finite set containing at least one point of infinite discontinuity, and no such point being on the frontier of the set of rectangles; and let D_n denote the remaining part of the fundamental rectangle when the finite set of rectangles is removed. Then, if $f(x, y)$ have a proper integral in every such domain D_n, and if this proper integral converge to a definite limit when any sequence whatever of such domains D_n is taken, such that the measure of D_n converges to that of the fundamental rectangle, this limit shall define the improper double integral of $f(x, y)$.

This extension of Cauchy's definition would admit of the existence of non-absolutely convergent improper double integrals, as in the case of improper single integrals. With this definition, the theorems of § 367 would no longer be valid.

When it is asserted that non-absolutely convergent double integrals do not exist, the assertion must be taken to mean that such integrals do not exist in accordance with the definition of Jordan, and not that it is impossible to give definitions, such as the above extension of Harnack, in accordance with which double integrals would exist that do not converge absolutely.

The properties of improper double integrals which are not necessarily absolutely convergent are more restricted than those which exist in accordance with Jordan's definition, and it is consequently a matter of opinion whether, though the former certainly exist as limits, the name integral may be appropriately applied to such limits.

EXAMPLE

If we take, as the domain of integration, the rectangle bounded by $x = 0$, $x = a$, $y = 0$, $y = b$, then the double integral of $\frac{1}{x}\sin\frac{1}{x}$ is not convergent, and therefore in accordance with Jordan's definition does not exist; although the single integral $\int_0^a \frac{1}{x}\sin\frac{1}{x}\,dx$ is non-absolutely convergent, and exists in accordance with Cauchy's definition.

The existence of $\int_0^a \frac{1}{x}\sin\frac{1}{x}\,dx$ depends upon the fact that $\int_\epsilon^a \frac{1}{x}\sin\frac{1}{x}\,dx$ converges to a definite limit, as ϵ converges to zero, and this is sufficient to ensure the existence of the single integral. Although $\int \frac{1}{x}\sin\frac{1}{x}\,d(x, y)$, taken over the domain bounded by $x = \epsilon$, $x = a$, $y = 0$, $y = b$, converges to a definite limit, as ϵ converges to zero, this is not sufficient to ensure the existence of the Jordan double integral. Taking Jordan's definition, let the domain D_n consist of the rectangular spaces bounded by the lines $y = 0$, $y = b$, and the lines parallel to the y-axis at the extremities of the intervals on the x-axis

$$\left(\frac{1}{(2n+1)\pi}, a\right), \quad \left(\frac{1}{(2n+3)\pi}, \frac{1}{(2n+2)\pi}\right), \quad \left(\frac{1}{(2n+5)\pi}, \frac{1}{(2n+4)\pi}\right), \ldots$$
$$\left(\frac{1}{(4n+1)\pi}, \frac{1}{4n\pi}\right).$$

The double integral taken through these spaces is

$$\int_{\frac{1}{(2n+1)\pi}}^{a} \int_0^b \frac{1}{x}\sin\frac{1}{x}\,dx\,dy + \sum_{p=n+1}^{p=2n} \int_{\frac{1}{(2p+1)\pi}}^{\frac{1}{2p\pi}} \int_0^b \frac{1}{x}\sin\frac{1}{x}\,dx\,dy,$$

or

$$\int_{\frac{1}{(2n+1)\pi}}^{a} \int_0^b \frac{1}{x}\sin\frac{1}{x}\,dx\,dy + b\int_0^\pi \sin z \sum_{p=n+1}^{p=2n} \frac{1}{z+2p\pi}\,dz;$$

which is greater than

$$b\int_{\frac{1}{(2n+1)\pi}}^{a} \frac{1}{x}\sin\frac{1}{x}\,dx + b\frac{n}{(4n+1)\pi}\int_0^\pi \sin z\,dz,$$

or than

$$b\int_{\frac{1}{(2n+1)\pi}}^{a} \frac{1}{x}\sin\frac{1}{x}\,dx + \frac{2nb}{(4n+1)\pi};$$

and this converges to

$$b\int_0^a \frac{1}{x}\sin\frac{1}{x}\,dx + \frac{b}{2\pi}:$$

whereas $\int \frac{1}{x}\sin\frac{1}{x}\,d(x, y)$, taken over the domain bounded by $x = \epsilon$, $x = a$, $y = 0$, $y = b$, converges to

$$b\int_0^a \frac{1}{x}\sin\frac{1}{x}\,dx.$$

Therefore the mode of choice of the intervals D_n affects the limit to which

$$\int_{D_n} \frac{1}{x}\sin\frac{1}{x}\,d(x, y)$$

converges, as D_n converges to the complete domain. Thus it is clear that the double integral, in accordance with Jordan's definition, does not exist.

THE DOUBLE INTEGRAL OVER AN INFINITE DOMAIN

369. The definition of a double integral has been extended by Jordan* to the case in which the field is unbounded. Let the function $f(x, y)$ be defined at every point of an unbounded set of points G, which is the outer limiting set of a sequence of closed domains, each of which is measurable (J); and let it be assumed that the upper and lower integrals of f exist over each bounded set of points Δ, that is measurable (J), contained in G, in accordance with the definition in § 338, or in accordance with that of § 366. This does not make it necessary that the function f should be bounded in G, even if it be bounded in every such set Δ. We shall denote the upper and lower integrals of f, in Δ, by $\bar{I}_\Delta(f)$, $\underline{I}_\Delta(f)$.

Consider a sequence $\{\Delta_n\}$, of sets Δ, such that Δ_n contains as a part, all those points of G, whose distance from the origin is $< \rho_n$: where $\{\rho_n\}$ is a monotone increasing sequence of positive numbers which increase indefinitely with n.

If $\bar{I}_{\Delta_n}(f)$ converges to a finite number $n \sim \infty$, independent of the particular sequence $\{\Delta_n\}$, subject to the above condition, then the upper integral $\bar{I}_G(f)$, of the function over the set G, is said to exist, and is defined by that number.

A similar definition may be applied to $\underline{I}_G(f)$, as the limit of $\underline{I}_{\Delta_n}(f)$.

When $\bar{I}_G(f)$, $\underline{I}_G(f)$ both exist, and are equal, their value is said to define $I_G(f)$, the integral of f over the unbounded domain G.

The necessary and sufficient condition for the existence of $\bar{I}_G(f)$ is that, if ϵ be an arbitrarily chosen positive number, the inequality $|\bar{I}_\Delta(f)| < \epsilon$ holds for every bounded set Δ, measurable (J), of points contained in G, and such that all the points of Δ are at a distance from the origin $\geqq \rho(\epsilon)$, where $\rho(\epsilon)$ is some positive number dependent on ϵ.

To prove that the condition is sufficient, let $\Delta^{(1)}$, $\Delta^{(2)}$ be any two bounded sets of points, measurable (J), each of them containing every point of G whose distance from the origin is $< \rho(\epsilon)$. Let $\delta^{(1)}$ be the set of points of $\Delta^{(1)}$ which do not belong to $\Delta^{(2)}$, and let $\delta^{(2)}$ be the set of points of $\Delta^{(2)}$ which do not belong to $\Delta^{(1)}$; thus the sets $\Delta^{(1)} + \delta^{(2)}$, $\Delta^{(2)} + \delta^{(1)}$ are identical. Therefore

$$I_{\Delta^{(1)}}(f) - \bar{I}_{\Delta^{(2)}}(f) = \bar{I}_{\delta^{(1)}}(f) - \bar{I}_{\delta^{(2)}}(f);$$

and since δ, δ' each contains no points of G at a distance $< \rho(\epsilon)$ from the origin, we have $|\bar{I}_{\Delta^{(1)}}(f) - \bar{I}_{\Delta^{(2)}}(f)| < 2\epsilon$.

If $\{\Delta_n\}$ be a sequence such as has been defined above, then if n be sufficiently large, $\rho_n > \rho(\epsilon)$; if then $\Delta^{(1)} = \Delta_n$, $\Delta^{(2)} = \Delta_{n+m}$, we have $|\bar{I}_{\Delta_n}(f) - \bar{I}_{\Delta_{n+m}}(f)| < 2\epsilon$, for a sufficiently large value of n, and for all positive values of m. Therefore $\{\bar{I}_{\Delta_n}(f)\}$ is a convergent sequence. Again, if $\{\Delta_n'\}$ be any other such sequence as $\{\Delta_n\}$, and n, n' are sufficiently large, we may take $\Delta^{(1)} = \Delta_n$, $\Delta^{(2)} = \Delta'_{n'}$; and thus

---

* See *Cours d'Analyse*, vol II, 2nd ed. p. 81.

$$| \bar{I}_{\Delta_n}(f) - \bar{I}_{\Delta'_{n'}}(f) | < 2\epsilon.$$

As this holds for every value of ϵ, when n, n' are sufficiently large, corresponding to each value of ϵ, the two sequences $\{\bar{I}_{\Delta_n}(f)\}$, $\{\bar{I}_{\Delta'_{n'}}(f)\}$ converge to the same limit; and thus $\bar{I}_G(f)$ exists as a finite number.

To shew that the condition is necessary, let us assume that it is not satisfied. Then ϵ can be so determined that, for every value of ρ, however large, there are sets Δ which contain no points of G at a distance $< \rho$ from the origin, and for which $|\bar{I}_\Delta(f)| \geq \epsilon$. If $\{\Delta_n\}$ be any sequence of the kind employed above, then, for any fixed value of n, the set Δ can be so determined that every point of it is at a distance from the origin greater than the upper boundary of the distances of all points of Δ_n; we have then $|\bar{I}_{\Delta_n+\Delta}(f) - \bar{I}_{\Delta_n}(f)| \geqq \epsilon$. If this be done for each value of n, the sequence $\{\bar{I}_{\Delta_n+\Delta}(f)\}$ cannot converge to the same limit as $\{\bar{I}_{\Delta_n}(f)\}$; the set Δ depends on n. It should be observed that the choice of Δ, for each value of n, is a case where the multiplicative axiom is employed.

The theorem for the lower integral may be proved in the same manner, or may be deduced by changing the sign of f.

370. The following theorem will be established:

The necessary and sufficient condition for the existence of the upper and lower integrals of f over the set G is that the upper integral of $|f|$ over G should exist as a finite number.

It will first be shewn that the condition is sufficient.

The upper boundary of $|f|$ in any cell is the greater of the numbers $|U|$, $|L|$; where U and L are the upper and lower boundaries of f in the cell, and the lower boundary is the lesser of these numbers. Either U or L may be infinite. If D be any set of points, measurable (J), and contained in Δ, we have accordingly

$$\bar{I}_D(|f|) \geqq |\bar{I}_D(f)|, \quad \bar{I}_D(|f|) \geqq |\underline{I}_D(f)|.$$

If $m(D)$ converge to the inner extent of Δ, we see that these relations still hold, when Δ is written for D.

If then $\bar{I}_\Delta(|f|) < \epsilon$, for every set of points Δ, all the points of which are at a distance $\geqq \rho(\epsilon)$ from the origin, we see that $|\bar{I}_\Delta(f)| < \epsilon$, and $|\underline{I}_\Delta(f)| < \epsilon$. Hence the conditions for the existence of $\bar{I}_G(f)$, $\underline{I}_G(f)$ are satisfied.

To prove the necessity of the condition, let $f = f^+ - f^-$, where $f^+ = f$, at every point at which f is $\geqq 0$, and elsewhere let $f^+ = 0$. By hypothesis, if Δ be any set of points such that all the points of G contained in it are at a distance $\geqq \sigma$ from the origin, we have $|\overline{I}_\Delta(f)| < \frac{1}{4}\epsilon$, $|\underline{I}_\Delta(f)| < \frac{1}{4}\epsilon$, provided σ is properly chosen.

Let Δ_1, Δ_2 be the two parts of Δ in which $f \geqq 0$, $f < 0$, respectively; and let Δ_1, Δ_2 contain sets D_1, D_2, measurable (J). $\overline{I}_\Delta(f^-)$ is the limit of $\overline{I}_{D_1}(f^+)$, or of $\overline{I}_{D_1+D_2}(f^+)$, as $m(D_1 + D_2)$ converges to the lower extent of Δ. Hence D_1, D_2 can be so determined that $\overline{I}_\Delta(f^-) - \overline{I}_{D_1+D_2}(f^-) < \eta$, and $\overline{I}_\Delta(f^-) - \overline{I}_{D_1+D_2}(f^-) < \eta$; where η is an arbitrarily chosen positive number.

Since the conditions $|\overline{I}_{D_1}(f)| < \frac{1}{4}\epsilon$, $|\underline{I}_{D_2}(f)| < \frac{1}{4}\epsilon$, hold by hypothesis, we have $\overline{I}_{D_1+D_2}(f^+) < \frac{1}{4}\epsilon$, $\overline{I}_{D_1+D_2}(f^-) < \frac{1}{4}\epsilon$, and therefore

$$\overline{I}_\Delta(f^+) < \tfrac{1}{4}\epsilon + \eta,\quad \overline{I}_\Delta(f^-) < \tfrac{1}{4}\epsilon - \eta.$$

Now $\overline{I}_\Delta(f^+ + f^-) = \lim \overline{I}_{D_1+D_2}(f^+ + f^-) = \overline{I}_\Delta(f^-) + \overline{I}_\Delta(f^-) \leqq \frac{1}{2}\epsilon - 2\eta$; and since η is arbitrary, $\overline{I}_\Delta(|f|) \leqq \frac{1}{2}\epsilon < \epsilon$; therefore $|f|$ has an upper integral over G. Thus the sufficiency of the condition has been established.

371. It will now be shewn that:

In order that $I_G(f)$ may exist, it is necessary and sufficient that $I_G(|f|)$ should exist as a definite number.

To shew that the condition is sufficient, we observe that, D being any set of points measurable (J), $|\overline{I}_D(f)|$ and $|\underline{I}_D(f)|$ both lie between $\overline{I}_D(|f|)$ and $\underline{I}_D(|f|)$. Hence we infer that, in the interval, $\overline{I}_G(|f|)$, $\underline{I}_G(|f|)$, both $\overline{I}_G(f)$, $\underline{I}_G(f)$ are contained. If then $\overline{I}_G(|f|)$ and $\underline{I}_G(|f|)$ are equal, which is the case when the integral $I_G(|f|)$ exists, it follows that $I_G(f)$ exists. To shew that the condition is necessary: we observe that if $I_D(f)$ exists, in any set D that is measurable (J), so also does $I_D(|f|)$ (see § 367), hence the same holds good for any set Δ, to which D converges; and proceeding to the limit, as Δ converges to G, we see that, if $I_G(f)$ exists, so also does $I_G(|f|)$, for $\overline{I}_G(|f|)$, $\underline{I}_G(|f|)$ both exist (§ 370), and each is the limit of $I_\Delta(|f|)$. It thus appears that, in accordance with Jordan's definition of the integral of $f(x, y)$ over an unbounded domain G, given in § 369, the integral is necessarily absolutely convergent, in the sense that $|f(x, y)|$ has a finite integral in the domain G. It has been seen, in § 358, that this is not the case for a single integral of a function $f(x)$ over an indefinitely great interval of x, as this integral may exist, and yet the integral of $|f(x)|$ over the same interval may not exist as a finite number. For example, the integral of $\sin x/x$ over the unbounded interval $(0, \infty)$ is not absolutely convergent. The reason for this difference between the two cases is that Jordan's definition for the double integral is much more stringent than the definition in § 354 for the single integral. The single integral is defined as the limit of a sequence of integrals taken each over a single finite interval. If the integrals, of which the integral over the unbounded interval is the limit, were not restricted in this manner, but might be integrals over other sets of points measurable (J), it would be possible to give a definition in accordance with which only absolutely convergent single integrals over an unbounded interval would exist. It would be possible to restrict the domains Δ in Jordan's definition in such a manner as to admit of the existence of non-absolutely convergent double integrals over an unbounded domain G.

When it is asserted that non-absolutely convergent double integrals over unbounded domains do not exist, the assertion must be taken to mean that such integrals do not exist in accordance with Jordan's definition, and not that it is impossible to give a definition in accordance with which double integrals exist that do not converge absolutely.

EXAMPLES

1*. The integral $$\int \sin(ax + by)\, x^{r-1} y^{s-1}\, d(x, y),$$
where $0 < r < 1$, $0 < s < 1$, taken over the positive quadrant, has no existence as an absolutely convergent improper integral. We find that the integral taken over the rectangle bounded by $x = 0$, $x = h$, $y = 0$, $y = k$ tends to the limit $a^{-r} b^{-s} \Gamma(r) \Gamma(s) \sin \frac{1}{2}(r + s)\pi$, as h and k are increased indefinitely. If the integral be taken over the domain $x > 0$, $y > 0$, $ax + by < h$, then when h is indefinitely increased, the integral has no limit if $1 < r + s < 2$; but it tends to the same limit as before, when $r + s < 1$. The integral may be regarded as conditionally convergent, if we adopt a definition in accordance with which it is sufficient that the integral taken through the rectangle $x = 0$, $x = h$, $y = 0$, $y = k$ should have a definite double limit, as h and k are indefinitely increased.

2†. The integrals
$$\int \cos(ax^2 + 2hxy + by^2)\, d(x, y), \qquad \int \sin(ax^2 + 2hxy + by^2)\, d(x, y),$$
where a, $ab - h^2$ are positive, taken over the positive quadrant, do not exist as absolutely convergent integrals. It may be seen that, if the integrals are taken over the quadrant of a circle bounded by $r = R$, the value of the integral has no definite limit, as R is increased indefinitely. If the integral be taken over the rectangle bounded by $x = 0$, $x = h'$, $y = 0$, $y = k'$, then, when h' and k' are increased indefinitely, the integrals have
$$0, \text{ and } \frac{1}{2\sqrt{ab - h^2}} \cos^{-1} \frac{h}{\sqrt{ab}}$$
for limits respectively, the inverse cosine having its least positive value. These may be regarded as the values of the integrals, subject to a suitable restriction on the domains of which the positive quadrant is the limit.

If $a = 0$, $b = 0$, $h = \frac{1}{2}$, $$\int \sin xy\, d(x, y),$$
over the positive quadrant, has no existence, even considered as the limit of an integral over the rectangle. But
$$\int \cos xy\, d(x, y)$$

---

* Hardy, *Messenger of Math.* vol. XXXII (1903), p. 96.

† *Ibid.* p. 159.

exists, and is equal to $\frac{1}{2}\pi$, when the integral is defined as the limit of the integral over the finite rectangle. It may be remarked that the single integrals

$$\int_0^{h'} \cos xy\,dx, \quad \int_0^{k'} \sin xy\,dy$$

are both non-convergent.

THE TRANSFORMATION OF DOUBLE INTEGRALS

372. Let (x, y) be a point of a bounded perfect and connex domain H, and let x and y be expressed by means of two functions f_1, f_2 in terms of two new variables ξ, η, which may be represented by points (ξ, η) in another plane. Let us suppose that the functions

$$x = f_1(\xi, \eta), \quad y = f_2(\xi, \eta),$$

and the reciprocal functions

$$\xi = \phi_1(x, y), \quad \eta = \phi_2(x, y),$$

are such that the following conditions are satisfied:

(1). To each point (x, y) there corresponds one point (ξ, η), and, conversely, to each point (ξ, η) there corresponds one point (x, y), and to the bounded domain H there corresponds a bounded domain $\overline{H}$.

(2) The functions $f_1(\xi, \eta)$, $f_2(\xi, \eta)$ are continuous functions of (ξ, η) throughout the domain $\overline{H}$.

(3). The functions $f_1(\xi, \eta)$, $f_2(\xi, \eta)$ have, at every point (ξ, η), of $\overline{H}$, definite partial differential coefficients with respect to ξ and η, and each one of these is everywhere continuous with respect to (ξ, η) and nowhere vanishes.

(4). The Jacobian of $f_1(\xi, \eta)$, $f_2(\xi, \eta)$ with respect to ξ and η does not vanish in the domain $\overline{H}$. In virtue of (3) the Jacobian is everywhere continuous, and of fixed sign.

From (2) and (3) it follows that, if $(\xi + \Delta\xi, \eta + \Delta\eta)$, (ξ, η) be two points of $\overline{H}$, and $(x + \Delta x, y + \Delta y)$, (x, y) the corresponding points of H, then

$$\Delta x = \left(\frac{\partial f_1}{\partial \xi} + \theta_1\right)\Delta\xi + \left(\frac{\partial f_1}{\partial \eta} + \theta_2\right)\Delta\eta,$$

$$\Delta y = \left(\frac{\partial f_2}{\partial \xi} + \theta_3\right)\Delta\xi + \left(\frac{\partial f_2}{\partial \eta} + \theta_4\right)\Delta\eta,$$

where $\theta_1, \theta_2, \theta_3, \theta_4$ converge to zero, as $\Delta\xi, \Delta\eta$ do so, and (see § 310) uniformly for all points (ξ, η) in any closed domain contained within $\overline{H}$. On solving these equations, we find

$$\Delta\xi = \frac{\left(\frac{\partial f_2}{\partial \eta} + \theta_4\right)\Delta x - \left(\frac{\partial f_1}{\partial \eta} + \theta_2\right)\Delta y}{J + \alpha_1},$$

with a similar expression for $\Delta\eta$, where J denotes the Jacobian

$$\frac{\partial(f_1, f_2)}{\partial(\xi, \eta)},$$

and α_1 is a function of $\theta_1, \theta_2, \theta_3, \theta_4$ which converges with them to zero. Since, by (4), J never vanishes, it follows from these equations that $\Delta\xi$, $\Delta\eta$ converge to zero with Δx, Δy, and thus that the functions $\phi_1(x, y)$, $\phi_2(x, y)$ are both continuous functions.

The partial differential coefficients

$$\frac{\partial\phi_1}{\partial x} = \frac{\partial f_2}{\partial\eta}\Big/J, \quad \frac{\partial\phi_1}{\partial y}, \quad \frac{\partial\phi_2}{\partial x}, \quad \frac{\partial\phi_2}{\partial y}$$

are also continuous in H; and therefore

$$\Delta\xi = \left(\frac{\partial\phi_1}{\partial x} + \chi_1\right)\Delta x + \left(\frac{\partial\phi_1}{\partial y} + \chi_2\right)\Delta y,$$

$$\Delta\eta = \left(\frac{\partial\phi_2}{\partial x} + \chi_3\right)\Delta x + \left(\frac{\partial\phi_2}{\partial y} + \chi_4\right)\Delta y,$$

where $\chi_1, \chi_2, \chi_3, \chi_4$ converge to zero with Δx, Δy, and uniformly so for all points (x, y) in a closed domain contained within H.

Corresponding to any closed set h, of zero content, contained within H. there is a closed set $\bar{h}$, of zero content, contained within $\bar{H}$. It is clear, from the continuity of the functions which define the transformation, that a limiting point of a sequence of points in H corresponds to the limiting point of the corresponding sequence in $\bar{H}$; and thus $\bar{h}$ is closed, since h is so.

Writing $\quad \Delta\xi = L\Delta x + M\Delta y, \ \ \Delta\eta = L'\Delta x + M'\Delta y,$

it follows, since

$$\frac{(\Delta\xi)^2 + (\Delta\eta)^2}{(\Delta x)^2 + (\Delta y)^2} \leq L^2 + L'^2 + M^2 + M'^2 + |LM + L'M'|,$$

where $L, L', \ldots$ converge uniformly to

$$\frac{\partial\phi_1}{\partial x}, \quad \frac{\partial\phi_2}{\partial x}, \quad \ldots,$$

that, if $|\Delta x|$, $|\Delta y|$ be both restricted to be less than a fixed positive number ϵ, the ratio

$$\frac{(\Delta\xi)^2 + (\Delta\eta)^2}{(\Delta x)^2 + (\Delta y)^2}$$

has a finite upper limit Λ^2, for the whole domain H. Now let the points of h be enclosed in a finite number of circles, the radii of which are all $< \epsilon$; it then follows that the points of $\bar{h}$ can be enclosed in a finite number of circles of which the radii are all less than $\epsilon\Lambda$. The sum of the areas of these circles on the (ξ, η) plane, which contain within them all the points of $\bar{h}$, has to the sum of the areas of the circles on the (x, y) plane, which enclose all the points of h, a ratio less than Λ^2. Since the sum of the latter circles can be taken to be arbitrarily small, it follows that the points of $\bar{h}$ can all be enclosed in a finite number of circles the sum of whose areas is arbitrarily small. Therefore $\bar{h}$ has the content zero.

373. Let $f(x, y)$ be a bounded function, defined for all points of a closed connex domain G, contained in H, the frontier of G having content zero; and let $f(x, y)$ be integrable (R) in G. If x, y be expressed in terms of ξ, η by the relations

$$x = f_1(\xi, \eta), \quad y = f_2(\xi, \eta),$$

which satisfy the conditions of § 372, then, corresponding to $f(x, y)$ in G, we have a function $F(\xi, \eta)$ in the domain $\bar{G}$, contained in $\bar{H}$, which corresponds to G. The frontier of $\bar{G}$, corresponding to the frontier of G, has also the content zero. A point of discontinuity of $f(x, y)$, in the (x, y) plane, corresponds to a point of discontinuity of $F(\xi, \eta)$ in the (ξ, η) plane, the measures of discontinuity at the corresponding points being the same. Since those points of (x, y) at which the saltus of $f(x, y)$ is $\geqq k$ form a closed set of zero content, it follows that the points of (ξ, η) at which the saltus of $F(\xi, \eta)$ is $\geqq k$ form also a closed set of zero content, and therefore $F(\xi, \eta)$ is integrable in $\bar{G}$.

In order to transform $\int f(x, y)\, d(x, y)$, taken throughout G, into an integral taken throughout $\bar{G}$, it is convenient to make use of an intermediate transformation* $x = \psi(u_1, u_2)$, $y = u_2$, followed by the transformation $u_1 = \xi$, $u_2 = f_2(\xi, \eta)$; the function $\psi(u_1, u_2)$ being such that

$$\psi(u_1, u_2) = f_1(\xi, \eta).$$

It is easy to see that each of these transformations satisfies the conditions of § 372.

Since $\dfrac{\partial x}{\partial u_1}$ is the Jacobian of (x, y) with respect to (u_1, u_2), we have, by a known theorem,

$$J = \frac{\partial x}{\partial u_1} \cdot \frac{\partial(u_1, u_2)}{\partial(\xi, \eta)};$$

hence, since J never vanishes,

$$\frac{\partial x}{\partial u_1} \text{ and } \frac{\partial(u_1, u_2)}{\partial(\xi, \eta)}$$

also never vanish.

---

* This method is employed in the general case of multiple integrals by Pierpont; see his paper "On multiple integrals," *Trans. Amer. Math. Soc.* vol. VI (1905), p. 416. It is, however, there assumed that $f(x, y)$ is integrable with respect to x for each value of y: but this is unnecessary.

Since $f(x, y)$ is integrable in G, we may, in accordance with the result of § 363, replace the double integral

$$\int f(x, y)\, d(x, y)$$

by the repeated integral $\displaystyle\int dy \overline{\int} f(x, y)\, dx,$

or by $\displaystyle\int dy \underline{\int} f(x, y)\, dx.$

Applying the transformation $x = \psi(u_1, u_2)$ to the upper and lower integrals

$$\overline{\int} f(x, y)\, dx, \quad \underline{\int} f(x, y)\, dx;$$

these may, in virtue of the theorem of § 360, be transformed into the single upper and lower integrals

$$\overline{\int} \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, du_1, \quad \underline{\int} \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, du_1,$$

where $\phi(u_1, u_2)$ represents the function of u_1, u_2 which corresponds to $f(x, y)$. We thus have

$$\begin{aligned} \int f(x, y)\, d(x, y) &= \int du_2 \overline{\int} \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, du_1 \\ &= \int du_2 \underline{\int} \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, du_1 \\ &= \int \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, d(u_1, u_2), \end{aligned}$$

the double integral being taken through the domain in the plane (u_1, u_2), which corresponds to G in (x, y).

Applying the same method of transformation to

$$\int \phi(u_1, u_2) \frac{\partial x}{\partial u_1}\, d(u_1, u_2),$$

where $\qquad u_1 = \xi, \quad u_2 = f_2(\xi, \eta),$

we have $\qquad \displaystyle\int f(x, y)\, d(x, y) = \int F(\xi, \eta) \frac{\partial x}{\partial u_1} \frac{\partial u_2}{\partial \eta}\, d(\xi, \eta),$

where $\qquad \displaystyle\frac{\partial u_2}{\partial \eta} = \frac{\partial(u_1, u_2)}{\partial(\xi, \eta)};$

hence finally we obtain the formula

$$\int f(x, y)\, d(x, y) = \int F(\xi, \eta)\, J d(\xi, \eta);$$

which is the formula of transformation of the integral of $f(x, y)$ throughout G into an integral throughout $\overline{G}$.

It has been assumed that J has a fixed sign throughout the domain of integration. If now this sign be negative, the product $\Delta\xi\Delta\eta$, in $J\Delta\xi\Delta\eta$, which corresponds to $\Delta x\Delta y$ in the plane of (x, y), must be accounted negative, when $\Delta x\Delta y$ is positive. It is however more convenient to consider $\Delta\xi\Delta\eta$ as essentially positive, otherwise the measure of a set of points in the (ξ, η) plane would have to be reckoned as negative. Adopting this convention, we write $|J| \, d(\xi, \eta)$ instead of $Jd(\xi, \eta)$: and therefore the formula of transformation will be written in the form

$$\int f(x, y) \, d(x, y) = \int F(\xi, \eta) \, |J| \, d(\xi, \eta).$$

374. Let us now assume that, at certain points of G, which form a set L of zero measure, either (1), $f(x, y)$ has an infinite discontinuity, or (2), one or more of the partial differential coefficients

$$\frac{\partial f_1}{\partial\xi}, \quad \frac{\partial f_1}{\partial\eta}, \quad \frac{\partial f_2}{\partial\xi}, \quad \frac{\partial f_2}{\partial\eta}$$

is discontinuous, or (3), the Jacobian J vanishes. In case J be positive over a part of G, and negative over another part, it is convenient to divide the double integral into two portions, taken over these two parts of G respectively, and to transform these two portions separately. It will accordingly be assumed that J never actually changes its sign in the domain G, although it may vanish at the points of the part L, of G. We may denote by $\overline{L}$ the set of points on the (ξ, η) plane which corresponds to L. It will be assumed that $\overline{L}$ has zero content. It will be shewn that, if one of the two integrals

$$\int f(x, y) \, d(x, y),$$

taken over G, and $\quad \int F(\xi, \eta) \, |J| \, d(\xi, \eta),$

taken over $\overline{G}$, exists as an absolutely convergent improper integral, or as a proper integral, then the other one exists, and the two have the same value.

Let us assume that $\int_G f(x, y) \, d(x, y)$ exists: it will then be sufficient, in order to establish the existence of the other integral, and its equality with the first, to shew that, for any domain $\overline{G_1}$, contained in $\overline{G}$, and itself containing no point of $\overline{L}$, either in its interior or on its frontier (which frontier is to be taken to be of zero measure), the condition

$$\left| \int_G f(x, y) \, d(x, y) - \int_{\overline{G_1}} F(\xi, \eta) \, |J| \, d(\xi, \eta) \right| < \bar{\eta}$$

is satisfied, provided that $m(\overline{G}) - m(\overline{G_1})$ be less than some fixed finite number dependent on $\bar{\eta}$.

A domain g, interior to G, and containing, in its interior and on its frontier, no points of L, can be found such that

$$\left|\int_G f(x, y)\, d(x, y) - \int_g f(x, y)\, d(x, y)\right| < \epsilon.$$

If h be any domain contained in g, such that $m(g) - m(h)$ is sufficiently small, we have

$$\left|\int_g f(x, y)\, d(x, y) - \int_h f(x, y)\, d(x, y)\right| < \epsilon;$$

and therefore $\left|\int_G f(x, y)\, d(x, y) - \int_h f(x, y)\, d(x, y)\right| < 2\epsilon.$

Now let k be a domain interior to G, containing in its interior, and on its frontier, no points of L, and containing h, then

$$\left|\int_G f(x, y)\, d(x, y) - \int_k f(x, y)\, d(x, y)\right| < 2\epsilon.$$

For, let p denote the domain obtained by taking the two domains g and k together, then

$$\left|\int_G f(x, y)\, d(x, y) - \int_p f(x, y)\, d(x, y)\right| < \epsilon,$$

$$\left|\int_p f(x, y)\, d(x, y) - \int_k f(x, y)\, d(x, y)\right| < \epsilon,$$

and by combining these inequalities the result follows.

We have now

$$\int_{\overline{G_1}} F(\xi, \eta)\, |J|\, d(\xi, \eta) = \int_{\overline{U}} F(\xi, \eta)\, |J|\, d(\xi, \eta) - \int_{\overline{V}} F(\xi, \eta)\, |J|\, d(\xi, \eta),$$

where $\overline{U}$ is the domain formed by taking all points which belong to one or both of the domains $\overline{G_1}$ and $\overline{h}$; and $\overline{V}$ consists of those points which belong to $\overline{h}$ but not to $\overline{G_1}$.

Now $\overline{U}$ corresponds to a domain in the (x, y) plane which contains h, and which domain may be taken to be identical with k; therefore

$$\int_{\overline{U}} F(\xi, \eta)\, |J|\, d(\xi, \eta) \text{ differs from } \int_G f(x, y)\, d(x, y)$$

by a number numerically less than 2ϵ. Again

$$\left|\int_{\overline{V}} F(\xi, \eta)\, |J|\, d(\xi, \eta)\right| < \mu\{m(\overline{G}) - m(\overline{G_1})\},$$

where μ is the upper boundary of $|F(\xi, \eta) J|$ in the domain $\overline{V}$, obtained by removing from $\overline{h}$ those points which belong to $\overline{G_1}$.

We thus have

$$\left|\int_G f(x, y)\, d(x, y) - \int_{\overline{G_1}} F(\xi, \eta)\, |J|\, d(\xi, \eta)\right| < 2\epsilon + \mu\{m(\overline{G}) - m(\overline{G_1})\}.$$

Now let ϵ be so fixed that it is $< \frac{1}{4}\bar{\eta}$, then $\bar{h}$ is fixed, so that μ cannot exceed a fixed finite number μ_1. If then $\overline{G_1}$ be so chosen that

$$m(\overline{G}) - m(\overline{G_1}) < \frac{\bar{\eta}}{2\mu_1},$$

the inequality

$$\left| \int_G f(x, y)\, d(x, y) - \int_{\overline{G_1}} F(\xi, \eta) \mid J \mid d(\xi, \eta) \right| < \bar{\eta}$$

will be satisfied Therefore it follows that

$$\int_{\overline{G}} F(\xi, \eta) \mid J \mid d(\xi, \eta)$$

exists, and is equal to $\int_G f(x, y)\, d(x, y)$.

375. This method of transformation may be extended to the case in which one of the domains G, $\overline{G}$ is infinite, or to the case in which both are infinite. It can be shewn that, if either of the integrals

$$\int_G f(x, y)\, d(x, y), \qquad \int_{\overline{G}} F(\xi, \eta) \mid J \mid d(\xi, \eta)$$

exists, the definition of § 366 being applied when G or $\overline{G}$ is infinite, then the other integral also exists, and the two integrals are equal. The proof can be given by slightly modifying the procedure of § 372.

There may be a set of points of zero measure, in the domain G, such that the corresponding values of ξ, η are infinite, or such that one of them is infinite. This set now takes the place of the set L. Whether G be finite or infinite, the domain h, contained in G, may be so fixed as to exclude all points which correspond to infinite values of ξ or η The domain k including h, and containing no points which correspond to infinite values of ξ and η, may then be fixed as before, and will satisfy the condition

$$\left| \int_G f(x, y)\, d(x, y) - \int_k f(x, y)\, d(x, y) \right| < 2\epsilon,$$

it being assumed that the integral of $f(x, y)$ over G exists.

The domain $\bar{h}$ contains all points of $\overline{G}$ of which the distance from the origin is less than some number R depending on the domain $G - h$, which contains in its interior all points (x, y) that correspond to infinite values of ξ or η, or of both. The same statement holds for $\bar{k}$, which contains $\bar{h}$. When the finite domain $\overline{G_1}$ is such that the condition

$$\int_{\overline{V}} F(\xi, \eta) \mid J \mid d(\xi, \eta) < \tfrac{1}{2}\bar{\eta}$$

is satisfied (and, in order that this may be the case, $\overline{G_1}$ must certainly contain all points of $\overline{G}$ whose distance from the origin is less than some fixed number $R_1 \leqq R$), we have as before

$$\left| \int_G f(x, y)\, d(x, y) - \int_{\overline{G}_1} F(\xi, \eta)\, |J|\, d(\xi, \eta) \right| < \bar{\eta};$$

and as $\bar{\eta}$ is an arbitrarily fixed number, we thus see that

$$\int_G F(\xi, \eta)\, |J|\, d(\xi, \eta)$$

exists, and is equal to $$\int_G f(x, y)\, d(x, y).$$

THE RIEMANN-STIELTJES INTEGRAL

376. The notion of the integral of a bounded function $f(x)$, defined in the linear interval (a, b), with respect to another bounded function $\phi(x)$, defined in the same interval, is a generalization of the integral of a function $f(x)$, with respect to the variable x, that was first introduced* into Analysis by Stieltjes in connection with the theory of continued fractions. This integral, in a generalized form, has recently become of considerable importance; and consequently, an account of it, so far as it can be regarded as a generalization of the R-integral, will be given here.

If $(a, x_1, x_2, \dots x_{m-1}, b)$ be a net, fitted on to (a, b), and in which the breadth of the greatest mesh is d, let us consider the sum

$$S_d \equiv f(\xi_1)\{\phi(x_1) - \phi(a)\} + f(\xi_2)\{\phi(x_2) - \phi(x_1)\} + \dots \\ + f(\xi_m)\{\phi(b) - \phi(x_{m-1})\},$$

where $f(x)$, $\phi(x)$ are the bounded functions, defined in the interval (a, b), and $\xi_1, \xi_2, \dots \xi_m$ are points, assigned in any manner, which are in the closed meshes $(a, x_1), (x_1, x_2) \dots (x_{m-1}, b)$ respectively.

We may denote x_0 by a, and x_m by b.

If the functions $f(x)$, $\phi(x)$ be such that S_d converges to a definite number, as the number m of the meshes is increased indefinitely, subject to the condition that d converges to zero, and if this limit is independent of the mode of successive sub-division of the interval by the nets, and of the mode in which the sets of points $\xi_1, \xi_2, \dots \xi_m$ are assigned in the nets, $f(x)$ is said to have a *Stieltjes Integral* with respect to $\phi(x)$. Such integral is defined to be the limit of S_d, as $d \sim 0$, and it is denoted by

$$\int_a^b f(x)\, d\phi(x).$$

It will here be spoken of as a Stieltjes integral.

In the case in which $f(x)$ is continuous in (a, b), and $\phi(x)$ is monotone and bounded in the same interval, the existence of the integral was established by Stieltjes. This will be proved below as a particular case of a more general theorem.

---

* See *Annales de la Faculté des Sciences de Toulouse*, vol. VIII (1894), p. J 71.

It will be shewn that:

(1). *If* $\int_a^b f(x)\, d\phi(x)$ *exists as a Stieltjes integral, then* $\int_a^b \phi(x)\, df(x)$ *also exists; and that*

$$\int_a^b f(x)\, d\phi(x) + \int_a^b \phi(x)\, df(x) = f(b)\, \phi(b) - f(a)\, \phi(a).$$

(2). *If* $\phi(x) = \phi_1(x) - \phi_2(x)$, *and the bounded function* $f(x)$ *has Stieltjes integrals with respect to the bounded functions* $\phi_1(x)$, $\phi_2(x)$, *then*

$$\int_a^b f(x)\, d\phi(x)$$

also exists as a Stieltjes integral, and has the value

$$\int_a^b f(x)\, d\phi_1(x) - \int_a^b f(x)\, d\phi_2(x).$$

To prove (1), we see that $\sum_{r=1}^{r=m} \phi(\xi_r)\, \Delta_{x_{r-1}}^{x_r} f(x)$, where $\Delta_{x_{r-1}}^{x_r} f(x)$ denotes $f(x_r) - f(x_{r-1})$, may be written in the form

$$-\sum_{r=1}^{r=m+1} f(x_{r-1})\, \Delta_{\xi_{r-1}}^{\xi_r} \phi(\xi) + f(b)\, \phi(b) - f(a)\, \phi(a),$$

where ξ_0 denotes a, and ξ_{m+1} denotes b The points $a, \xi_1, \ldots, \xi_m, b$ define a net with $m+1$ meshes, the breadth of each of which is $\leq 2d$. If d be sufficiently small, since $\int_a^b f(x)\, d\phi(x)$ exists, we see that

$$\sum_{r=1}^{r=m+1} f(x_{r-1})\, \Delta_{\xi_{r-1}}^{\xi_r} \phi(\xi)$$

differs from this integral by less than ϵ: therefore $\sum_{r=1}^{r=m} \phi(\xi_r)\, \Delta_{x_{r-1}}^{x_r} f(x)$ differs from

$$-\int_a^b f(x)\, d\phi(x) + f(b)\, \phi(b) - f(a)\, \phi(a),$$

by less than ϵ, provided d is sufficiently small, however the points ξ are chosen. Since ϵ is arbitrary, it follows that $\int_a^b \phi(x)\, df(x)$ exists, and is equal to

$$-\int_a^b f(x)\, d\phi(x) + f(b)\, \phi(b) - f(a)\, \phi(a).$$

In order to prove (2), we employ the identity

$$\sum_{r=1}^{r=m} f(\xi_r)\, \Delta_{x_{r-1}}^{x_r} \phi_1(x) - \sum_{r=1}^{r=m} f(\xi_r)\, \Delta_{x_{r-1}}^{x_r} \phi_2(x) = \sum_{r=1}^{r=m} f(\xi_r)\, \Delta_{x_{r-1}}^{x_r} \phi(x),$$

and the result follows by taking the limits of the expressions on the left-hand side.

As in the case of the R-integral (see § 330), we may, in the definition given above of the Stieltjes integral, replace the numbers $f(\xi_r)$, where $r = 1, 2, 3 \dots m$, by any numbers $M_{x_{r-1}}^{x_r} f(x)$, where $M_{x_{r-1}}^{x_r} f(x)$ is restricted to be in the interval bounded by $U_{x_{r-1}}^{x_r} f(x)$, $L_{x_{r-1}}^{x_r} f(x)$; these numbers denoting the upper and lower boundaries of $f(x)$ in the interval (x_{r-1}, x_r); we thus obtain the following definition of the *Riemann-Stieltjes or RS-integral*:

If S_D denotes the sum $\sum_{r=1}^{r=m} M_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x)$ taken for a net

$$(a, x_1, x_2, \dots x_{m-1}, b),$$

fitted on to (a, b), the breadth of the greatest mesh being d, where $M_{x_{r-1}}^{x_r} f(x)$ is in the interval bounded by the upper and lower boundaries of $f(x)$ in the closed mesh (x_{r-1}, x_r); and if S_D converges to a definite number, as the number m is increased indefinitely, subject only to the condition that d converges to zero, the limit of S_D being assumed to be independent of the mode in which the successive nets are defined, and of the mode in which the numbers $M_{x_{r-1}}^{x_r} f(x)$ are chosen, then the limit of S_D is said to define the RS-integral $\int_a^b f(x)\, d\phi(x)$.

It is clear that, in case the RS-integral of $f(x)$ with respect to $\phi(x)$ exists, the Stieltjes integral also exists. It will be shewn that the converse also holds good. In the case $\phi(x) = x$, we have the ordinary R-integral; and in this case the equivalence of the two definitions has been established in § 330.

In accordance with the above definition of the RS-integral, when the integral exists, either of the sums

$$\sum_{r=1}^{r=m} U_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x), \quad \sum_{r=1}^{r=m} L_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

has the RS-integral for its limit. Assuming the existence of the Stieltjes integral, since it is always possible to determine the points $\xi_1, \xi_2, \dots \xi_m$ so that $0 \leq U_{x_{r-1}}^{x_r} f(x) - f(\xi_r) < \epsilon$, for $r = 1, 2, \dots m$; where ϵ is a prescribed positive number, we see that

$$\left| \sum_{r=1}^{r=m} U_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x) - \sum_{r=1}^{r=m} f(\xi_r)\, \Delta_{x_{r-1}}^{x_r} \phi(x) \right| < \epsilon \Sigma \left| \Delta_{x_{r-1}}^{x_r} \phi(x) \right|.$$

For the particular net D, ϵ may be so chosen that $\epsilon \left| \Delta_{x_{r-1}}^{x_r} \phi(x) \right|$ does not exceed an arbitrarily chosen number $\eta(D)$. Thus the number on the left-hand side is $< \eta(D)$. In case the function $\phi(x)$ is of bounded variation in (a, b), the number $\Sigma \left| \Delta_{x_{r-1}}^{x_r} \phi(x) \right|$ is less than a number K, independent of the net; in this case we may take ϵ to depend only upon d, the maximum mesh of D, and so that ϵ converges to zero, as $d \sim 0$. We then see that

$$\sum_{r=1}^{r=m} U_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

has the same limit as

$$\sum_{r=1}^{r=m} f(\xi_r)\, \Delta_{x_{r-1}}^{x_r} \phi(x),$$

which is the Stieltjes integral. In case $\phi(x)$ is not of bounded variation, we consider a particular sequence of nets such that d converges in the sequence to zero, whether the sequence forms a system of nets or not. The values of ϵ in the sequence may be so chosen that $\eta(D)$ converges to zero; thus the equality of the limits holds for any particular sequence: and we have the same result as before. In a similar manner it can be shewn that

$$\sum_{r=1}^{r=m} L_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

has the Stieltjes integral as its limit. It then follows that

$$\sum_{r=1}^{r=m} M_{x_{r-1}}^{x_r} f(x)\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

has the Stieltjes integral for its limit; and therefore the RS-integral exists.

The most important case is that in which $\phi(x)$ is of bounded variation in (a, b), or in particular is bounded and monotone. Since the RS-integral exists when the Stieltjes integral exists, in virtue of the theorem (2) above, the relation

$$\int_a^b f(x)\, d\phi(x) = \int_a^b f(x)\, d\phi_1(x) - \int_a^b f(x)\, d\phi_2(x)$$

holds good for RS-integrals, where $\phi(x) = \phi_1(x) - \phi_2(x)$, the functions $\phi_1(x)$, $\phi_2(x)$ being monotone, and it being assumed that the integrals on the right-hand side exist.

It can easily be shewn that the RS-integral cannot exist if, at any point $\bar{x}$ of (a, b), both the functions $f(x)$, $\phi(x)$ are discontinuous. For, let a net be so chosen that $\bar{x}$ is interior to the mesh (x_{r-1}, x_r): for this mesh we may take $M_{x_{r-1}}^{x_r} f(x)$ to have either of the values $U_{x_{r-1}}^{x_r} f(x)$, $L_{x_{r-1}}^{x_r} f(x)$: and the corresponding sums will differ by

$$(U_{x_{r-1}}^{x_r} f(x) - L_{x_{r-1}}^{x_r} f(x))\, \Delta_{x_{r-1}}^{x_r} \phi(x),$$

provided the values of $M_{x_{s-1}}^{x_s} f(x)$ for $s \neq r$ are taken to be the same in the two cases. This difference is greater than some fixed number, if the interval (x_{r-1}, x_r), including $\bar{x}$ in its interior, is properly chosen, when $f(x)$, $\phi(x)$ are both discontinuous at $\bar{x}$, the trivial case in which $\phi(x)$ has merely a removable discontinuity at $\bar{x}$ being excluded. It follows that a sequence of nets and two sets of values of $M_{x_{r-1}}^{x_r} f(x)$, which differ only in one mesh of each net, can be so chosen that the corresponding sums cannot converge to the same limit.

376¹. The following theorem* will be established for the RS-integral:

If $f(x)$ be bounded in (a, b), and $\phi(x)$ be of bounded variation in (a, b), the necessary and sufficient condition that $f(x)$ should have an RS-integral with respect to $\phi(x)$ is that the variation of $\phi(x)$ over the set of points of discontinuity of $f(x)$ should be zero.

It will be observed that the condition cannot be satisfied if $f(x)$, $\phi(x)$ have any common point of discontinuity.

The upper and lower variations of $\phi(x)$, in case it be monotone, over a set of points in (a, b), have been defined in § 252; when the upper variation is zero, the variation over the set of points exists, and is equal to zero.

We may, without loss of generality, assume $\phi(x)$ to be monotone and non-diminishing; for in the general case $\phi(x)$ is the difference of two such functions $P(x)$, $N(x)$, and the variation of $\phi(x)$ over a set of points is the sum of the variations of $P(x)$ and $N(x)$ over the set (see § 244). We assume accordingly that $\phi(x)$ is monotone non-diminishing in (a, b).

Let us consider a system of nets $\{D_n\}$, with closed meshes (see § 51) fitted on to (a, b); and let the sums $\overline{S}_{D_n}$, $\underline{S}_{D_n}$ corresponding to the nth net of the system be defined by

$$\overline{S}_{D_n} = \sum_{r=1}^{r=m_n} U_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x), \quad \underline{S}_{D_n} = \sum_{r=1}^{r=m_n} L_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x)$$

It is easily seen that, when a net (x_{r-1}, x_r) is divided into two parts, $U_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x)$ cannot be increased, and $L_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x)$ cannot be diminished. For, if x' be the point of division, $U_{x_{r-1}}^{x_r} f(x)\ \Delta_{x_{r-1}}^{x_r} \phi(x)$ becomes

$$U_{x_{r-1}}^{x'} f(x) . \Delta_{x_{r-1}}^{x'} \phi(x) + U_{x'}^{x_r} f(x)\ \Delta_{x'}^{x_r} \phi(x),$$

where
$$\Delta_{x_{r-1}}^{x_r} \phi(x) = \Delta_{x_{r-1}}^{x'} \phi(x) + \Delta_{x'}^{x_r} \phi(x),$$

and
$$U_{x_{r-1}}^{x'} f(x) \leqq U_{x_{r-1}}^{x_r} f(x), \quad U_{x'}^{x_r} f(x) \leqq U_{x_{r-1}}^{x_r} f(x),$$

from which the result follows at once. For $L_{x_{r-1}}^{x_r} f(x)\ \Delta_{x_{r-1}}^{x_r} \phi(x)$ the corresponding result is proved in a similar manner.

It follows that $\overline{S}_{D_n}$ cannot increase, and $\underline{S}_{D_n}$ cannot diminish as n is increased; thus $\{\overline{S}_{D_n}\}$ is a monotone non-increasing, and $\{\underline{S}_{D_n}\}$ a monotone non-diminishing, sequence, for the system of nets. Hence $\overline{S}_{D_n}$ has a lower boundary $\overline{S}$, and $\underline{S}_{D_n}$ an upper boundary $\underline{S}$, for the system of nets.

It has been shewn in § 334 that, if ζ be a prescribed positive number, the interval (a, b) can be covered by a finite set of intervals, each of length $\leqq \zeta$, such that, in each interval, the fluctuation of $f(x)$ is $< \omega(\xi) + \epsilon$, where ξ is a definite point interior to the interval, and $\omega(\xi)$ is the saltus at ξ, and ϵ is an arbitrarily chosen number. The end-points of this finite set of intervals divide (a, b) into a finite number of parts, let η be the length of the least of these parts. Any interval whatever δ, in (a, b), of length $< \eta$, is interior to one of the intervals of the finite set. Hence, in δ, the fluctuation of $f(x)$ is $< \epsilon + \omega(\xi)$, where ξ is some point in an interval of length $\leqq \zeta$, which contains δ in its interior.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol XIII (1913), p. 133.

Let G_ϵ be the closed set of points at which $\omega(x) \geqq \epsilon$, with centre x, any point of G_ϵ, an interval of length 2ρ can be taken: all such intervals, or the parts of them in (a, b), coalesce into a non-overlapping finite set $H_{\epsilon,\rho}$, of which G_ϵ is the inner limiting set, as $\rho \sim 0$. Assuming that the variation of $\phi(x)$ over G_ϵ is zero, ρ can be so chosen that the variation of $\phi(x)$ over $H_{\epsilon,\rho}$ is $< \zeta'$, an arbitrarily chosen positive number. In any interval of the set $H_{\epsilon,\rho}$ there is no point of G_ϵ of which the distance from either end-point is $< \rho$. We can suppose ζ, and consequently η, to be so small that, if δ be not interior to an interval of $H_{\epsilon,\rho}$, there is no point of G_ϵ in the interval δ.

It will now be proved that $\overline{S} = \underline{S}$, if the condition that the variation of $\phi(x)$ over the set of discontinuities of $f(x)$ is zero be satisfied. We have

$$\overline{S}_{D_n} - \underline{S}_{D_n} = \sum_{r=1}^{r=m_n} \{U_{x_r, x_{r-1}} f(x) - L_{x_r, x_{r-1}} f(x)\} \Delta_{x_r, x_{r-1}} \phi(x).$$

The number n being taken so large that every mesh of D_n is $< \eta$, we find that

$$\overline{S}_{D_n} - \underline{S}_{D_n} < 2\epsilon \{\phi(b) - \phi(a)\} + \zeta' (U - L),$$

where the first term on the right-hand side arises from those meshes that are not interior to an interval of $H_{\epsilon,\rho}$; and in each such mesh the fluctuation of $f(x)$ is $< 2\epsilon$. The second term arises from the remaining meshes. Since ϵ, ζ' can be made to converge to zero, as $n \sim \infty$, we see that $\overline{S} = \underline{S}$.

This has been shewn to hold good for any one system of nets. It will now be shewn that the number $S = \overline{S} = \underline{S}$ is the same for all systems of nets. If possible, let it be assumed that for a second system of nets, the number $S' = \lim_{n \sim \infty} \overline{S}_{D_n}' = \lim_{n \sim \infty} \underline{S}_{D_n}'$ has a value different from S, and let it be supposed that $S' > S$. If ϵ be an arbitrarily chosen positive number, n may be taken so large that

$$\overline{S}_{D_n} < S + \epsilon, \quad \underline{S}_{D_n} > S - \epsilon, \quad \overline{S}_{D_n}' < S' + \epsilon, \quad \underline{S}_{D_n}' > S' - \epsilon.$$

Let the two nets of order n in the two systems be superimposed, and let $\overline{S}_n''$, $\underline{S}_n''$ be the sums for the new net so obtained.

We have then

$$\overline{S}_{D_n}'' \leqq \overline{S}_{D_n} < S + \epsilon, \quad \underline{S}_{D_n}'' \geqq \underline{S}_{D_n} > S - \epsilon;$$

and similarly

$$\overline{S}_{D_n}'' < S' + \epsilon, \quad \underline{S}_{D_n}'' > S' - \epsilon;$$

from these inequalities we deduce that

$$\overline{S}_{D_n}'' - \underline{S}_{D_n}'' < S - S' + 2\epsilon.$$

If ϵ be chosen to be less than the positive number $\frac{1}{2}(S' - S)$, we see that $\overline{S}_{D_n}'' - \underline{S}_{D_n}''$ has a negative value, and this is impossible. It has thus been shewn that $S = S'$; therefore the number S is the same for all systems of nets.

It will next be shewn that the convergence of S_{D_n} to S, or of $\underline{S}_{D_n}$ to S, is uniform with respect to all possible systems of nets, in the sense that a number d can be so chosen that, for any net of which all the meshes are $\leq d$, the sum of the products of the fluctuation of $f(x)$ in a mesh into the difference of values of $\phi(x)$ at the ends of the mesh is less than a prescribed positive number ϵ.

Denoting $U_{x_{r-1}}^{x_r} f(x) - L_{x_{r-1}}^{x_r} f(x)$ by $F_{x_{r-1}}^{x_r} f(x)$, let a fixed net D_1 be so chosen that $\sum_{r=1}^{r=m} F_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x) < \frac{1}{2}\epsilon$.

We consider a second net D_2, for which all the meshes do not exceed in length a number d to be fixed hereafter, and which may be taken in the first instance so small that no mesh of D_2 contains more than one of the points x_r in its interior or at an end-point. Those meshes of D_2 that contain, as interior point, or as an end-point, one of the points x_r are at most $2m$ in number; the remaining meshes of D_2 being interior to meshes of D_1. A mesh δ of D_2 which contains x_r, either as an interior point or as an end-point, is such that the term corresponding to it in the sum for D_2 is the product of the fluctuation of $f(x)$ in δ ($\leq d$) into the difference of the functional values of $\phi(x)$ at the ends of δ. Since $f(x)$, $\phi(x)$ are not both discontinuous at x_r, one or other of the factors in this term diminishes indefinitely as d does so, the other factor being less than a fixed number. The number d can accordingly be taken so small that the magnitude of the term in the sum for D_2 does not exceed $\frac{\epsilon}{4m}$. The value of d can be chosen so small that this condition is satisfied for each mesh which contains one of the points x_r; then the sum of the corresponding terms in the sum for D_2 is $\leq \frac{1}{2}\epsilon$. The remaining part of the sum for D_2 is less than $\frac{1}{2}\epsilon$. Therefore the sum for D_2 is less than ϵ. Therefore d has been so chosen that, for any net D_2 whose meshes are all of length $\leq d$, the sum

$$\sum_{r=1}^{r=m'} F_{x'_{r-1}}^{x_r'} f(x) . \Delta_{x'_{r-1}}^{x_r'} \phi(x) \text{ is } < \epsilon,$$

which is what we had to prove. The sufficiency of the condition stated in the theorem has now been proved.

To shew that the condition that the variation of $\phi(x)$ over the set of points of discontinuity of $f(x)$ is necessary, in order that $\overline{S} = \underline{S}$, we observe that $\overline{S}_{D_n} - \underline{S}_{D_n}$ is $\geq \frac{1}{2}\epsilon . \Sigma \{\phi(x_r) - \phi(x_{r-1})\}$, where the summation is taken for those meshes of D_n which contain points of G_ϵ within them, or at an end-point. For, if a point of G_ϵ be a common end-point of two meshes, in one at least of those meshes the fluctuation of $f(x)$ is $\geq \frac{1}{2}\epsilon$. Unless the variation of $\phi(x)$ over G_ϵ is zero, the sum

$$\Sigma \{\phi(x_r) - \phi(x_{r-1})\}$$

remains greater than some positive number, however large n may be, and in that case $\overline{S}_{D_n} - \underline{S}_{D_n}$ cannot converge to zero. Thus it is necessary that the variation of $\phi(x)$ over G_ϵ should be zero for every value of ϵ, and this cannot be the case unless the variation of $\phi(x)$ over G, the outer limiting set of G_ϵ, for a sequence of diminishing values of ϵ, converging to zero, is also zero.

It should be observed that the condition that $\phi(x)$ has variation zero over the set of points of discontinuity of $f(x)$ may be satisfied even when the measure of the set is > 0. For example, if $\phi(x) = x$, at the points of a closed set E, of positive measure, and if $\phi(x)$ is constant over each contiguous interval of E, whilst all the discontinuities of $f(x)$ are in these contiguous intervals, $\int_a^b f(x)\,d\phi(x)$ exists as an RS-integral, although $\int_a^b f(x)\,dx$ may not exist as an R-integral.

376^2. It is seen as a particular case of the above general theorem that:

In case $f(x)$ is continuous, the RS-integral of $f(x)$ with respect to any bounded monotone function, or function of bounded variation, always exists. In case $\phi(x)$ is absolutely continuous, and $f(x)$ has an R-integral in (a, b), the RS-integral $\int_a^b f(x)\,d\phi(x)$ always exists.

We need only consider the case in which the absolutely continuous function $\phi(x)$ is monotone, for, in the general case, $\phi(x)$ is the difference of two such functions. The set of points of discontinuity of $f(x)$ has measure zero, and the variation of an absolutely continuous function over a set of measure zero is zero (see § 352). Thus the condition in the general theorem is satisfied.

In case $\phi(x)$ is continuous, and of bounded variation, and $f(x)$ has only an enumerable set of points of discontinuity, and in particular in case it is of bounded variation, the RS-integral of $f(x)$ with regard to $\phi(x)$ exists.

We need only consider the case in which $\phi(x)$ is monotone and continuous. To an enumerable set of points on the x-segment, there corresponds an enumerable set of points on the ξ-segment, where $\xi = \phi(x)$. The measure of this latter set is zero, and therefore the measure of $\phi(x)$ over the set of discontinuities of $f(x)$ is zero; the sufficient condition is thus satisfied.

If $f(x)$, $\phi(x)$ are both of bounded variation in (a, b) and have no points of discontinuity in common, either of them has an RS-integral with respect to the other.

We consider the case in which $\phi(x)$ is monotone, as the general case is deducible from this. Since $f(x)$ is of bounded variation, its set of points of discontinuity is enumerable. Since $\phi(x)$ is continuous at all the points of this set, its measure over the set is zero (see § 252); and this is the sufficient condition for the existence of $\int_a^b f(x)\, d\phi(x)$.

THE GENERALIZED RIEMANN-STIELTJES INTEGRAL

377. Definitions of an RS-integral have been given by Hardy[*], W. H. Young[†], and Pollard[‡]. These definitions vary in scope, and some of them are more general than the definition given in § 376.

If $f(x)$, $\phi(x)$ be any two functions, bounded in (a, b), and such that $\phi(x)$ is monotone non-diminishing; let D denote a net with closed meshes, fitted on to (a, b). We consider the two sums, for the net D,

$$\overline{S}_D = \sum_{r=1}^{r=m} U_{x_{r-1}}^{x_r} f(x) \cdot \Delta_{x_{r-1}}^{x_r} \phi(x), \quad \underline{S}_D = \sum_{r=1}^{r=m} L_{x_{r-1}}^{x_r} f(x) \cdot \Delta_{x_{r-1}}^{x_r} \phi(x).$$

It is clear that $\overline{S}_D \geqq L\{\phi(b) - \phi(a)\}$, $\underline{S}_D \leqq U\{\phi(b) - \phi(a)\}$; where U and L are the upper and lower boundaries of $f(x)$ in (a, b). It follows that, when all possible nets D are considered, $\overline{S}_D$ has a finite lower boundary, and $\underline{S}_D$ has a finite upper boundary; these may be denoted by

$$(G)\overline{\int_a^b} f(x)\, d\phi(x), \text{ or } \overline{I}, \quad (G)\underline{\int_a^b} f(x)\, d\phi(x), \text{ or } \underline{I},$$

respectively. They may be spoken of as the *upper and lower generalized RS-integrals* of $f(x)$ with respect to $\phi(x)$, in (a, b). If ϵ be an arbitrarily chosen positive number, it is possible to determine a net D, for which

$$\overline{S}_D < (G)\overline{\int_a^b} f(x)\, d\phi(x) + \epsilon,$$

and also a net D', for which

$$\underline{S}_{D'} > (G)\underline{\int_a^b} f(x)\, d\phi(x) - \epsilon.$$

It can be shewn that

$$(G)\overline{\int_a^b} f(x)\, d\phi(x) \geqq (G)\underline{\int_a^b} f(x)\, d\phi(x), \text{ or } \overline{I} \geqq \underline{I}.$$

---

[*] *Messenger of Math* vol. XLVIII (1918), p. 90.

[†] *Proc. Lond. Math. Soc.* (2), vol XIII (1913), p. 109. *ibid.* vol. XV (1916), p. 35.

[‡] *Quarterly Journal of Math.* vol. XLIX (1923), p. 73, where a discussion of several definitions is given, and the properties of the corresponding integrals are discussed in detail.

Assume that, if possible $\bar{I} < \underline{I}$, then for the net D, $\overline{S}_D < \bar{I} - \epsilon$: and for the net D', $\underline{S}_{D'} > \underline{I} - \epsilon$. If we superimpose the nets D, D' into a single net (D, D'), we have

$$\overline{S}_{(D,D')} \leqq S_D < \bar{I} + \epsilon, \quad \underline{S}_{(D,D')} \geqq \underline{S}_{D'} > \underline{I} - \epsilon.$$

Choosing ϵ to be less than $\frac{1}{2}(\underline{I} - \bar{I})$, we see, from these inequalities, that $\underline{S}_{(D,D')} > \overline{S}_{(D,D')}$, which is impossible in accordance with the definitions of $\underline{S}_{(D,D')}$, $\overline{S}_{(D,D')}$. Therefore $\bar{I}$ cannot be less than $\underline{I}$.

In case the upper and lower generalized RS-integrals $\bar{I}$, $\underline{I}$ are equal, their common value I is said to define the generalized RS-integral

$$(G) \int_a^b f(x)\, d\phi(x),$$

which then exists.

It will be shewn that:

When the generalized RS-integral exists, there exists a system of nets $\{D_n\}$, such that $\overline{S}_{D_n}$, $\underline{S}_{D_n}$ both converge to I, as $n \sim \infty$.

If $\{\epsilon_n\}$ be a sequence of diminishing numbers converging to zero, a set of nets $D_1^{(1)}, D_2^{(1)}, \ldots D_n^{(1)}, \ldots$, for which $d_n^{(1)}$, the maximum length of a net, converges to zero, as $n \sim \infty$, exists such that $\overline{S}_{D_n^{(1)}} < I - \epsilon_n$, for $n = 1, 2, 3, \ldots$; and there also exists a set of nets $D_1^{(2)}, D_2^{(2)}, \ldots D_n^{(2)}, \ldots$, for which $\underline{S}_{D_n^{(2)}} > I - \epsilon_n$, and for which $d_n^{(2)}$ converges to zero as $n \sim \infty$. Consider the sequence of nets

$(D_1^{(1)}, D_1^{(2)}), (D_1^{(1)}, D_2^{(1)}; D_1^{(2)}, D_2^{(2)}), \ldots, (D_1^{(1)}, D_2^{(1)}, \ldots D_n^{(1)}, D_1^{(2)}, \ldots, D_r^{(2)}), \ldots$. This sequence forms a system $\bar{D}$ of nets, and for the nth net $\bar{D}_n$ of this system we have $\overline{S}_{\bar{D}_n} < I + \epsilon_n$, $\underline{S}_{\bar{D}_n} > I - \epsilon_n$. It then follows that

$$\lim_{n \sim \infty} \overline{S}_{\bar{D}_n} = \lim_{n \sim \infty} \underline{S}_{\bar{D}_n} = I.$$

Consequently we have the following statement:

The necessary and sufficient condition that the generalized RS-integral of a bounded function $f(x)$ with respect to a monotone function $\phi(x)$ exists is that, a system $\bar{D}$ of nets exists such that

$$\sum_{r=1}^{r=m_n} F_{x_{r-1}}^{x_r} f(x) \cdot \Delta_{x_{r-1}}^{x_r} \phi(x),$$

taken for a net $\bar{D}_n$ of the system, converges to zero, as $n \sim \infty$. $F_{x_{r-1}}^{x_r} f(x)$ denotes the fluctuation of $f(x)$ in the interval (x_{r-1}, x_r).

It is not necessarily the case that, when I exists, the numbers $\overline{S}_{D_n}$, $\underline{S}_{D_n}$, for an arbitrarily chosen system of nets, converge to I, although, as has been shewn, it is possible to choose the system of nets so that this is the case. For any particular system of nets $\lim\limits_{n \sim \infty} \overline{S}_{D_n}$ must be $\geqq I$, and $\lim\limits_{n \sim \infty} \underline{S}_{D_n}$ must be $\leqq I$. Thus S_{D_n} and $\underline{S}_{D_n}$ cannot have the same limit, for $n \sim \infty$, unless that limit has the value I.

It may happen that the generalized *RS*-integral exists in a case in which $f(x)$, $\phi(x)$ have one or more points of discontinuity in common, and when consequently the *RS*-integral does not exist. It will however be shewn that, when the generalized *RS*-integral exists, such a common point of discontinuity of $f(x)$ and $\phi(x)$ must be an end-point of meshes of the special system $\bar{D}$ of nets. For, if possible, let a point ξ, which is, for every value of n, an interior point of a mesh d_n of the system $\bar{D}$, be a point at which $f(x)$ and $\phi(x)$ are both discontinuous. The difference $\bar{S}_{\bar{D}_n} - \underline{S}_{\bar{D}_n}$ contains a term $F_{d_n} f(x) . \Delta_{d_n} \phi(x)$, where $F_{d_n} f(x)$ is the fluctuation of $f(x)$ in d_n, and $\Delta_{d_n} \phi(x)$ is the difference of the values of $\phi(x)$ at the end-points of d_n. Since d_n contains the point ξ, both $F_{d_n} f(x)$ and $\Delta_{d_n} \phi(x)$ are, for all values of n, not less than fixed positive numbers; therefore $\bar{S}_{D_n} - \underline{S}_{D_n}$ cannot converge to zero, as $n \sim \infty$. Thus no such point as ξ can exist.

It will further be shewn that, when I exists, and when $f(x)$, $\phi(x)$ are both discontinuous at an end-point of meshes of the system $\bar{D}$, $f(x)$ must be continuous at the point on one side, and $\phi(x)$ must be continuous at the point on the other side*.

We have, for a mesh of the net $\bar{D}_n$

$$U_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x) \geqq U_{x_{r-1}}^{x_{r-1}+\epsilon} f(x) . \Delta_{x_{r-1}}^{x_{r-1}+\epsilon} \phi(x)$$
$$+ U_{x_{r-1}+\epsilon}^{x_r-\epsilon} f(x) . \Delta_{x_{r-1}+\epsilon}^{x_r-\epsilon} \phi(x) + U_{x_r-\epsilon}^{x_r} f(x) . \Delta_{x_r-\epsilon}^{x_r} \phi(x),$$

where ϵ is any positive number $< \frac{1}{2}(x_r - x_{r-1})$. Let ϵ converge to zero; we then have

$$U_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x) \geqq M_{x_{r-1}}^{+} f(x) [\phi(x_{r-1}+0) - \phi(x_{r-1})]$$
$$+ U_{x_{r-1}+0}^{x_r-0} f(x) [\phi(x_r - 0) - \phi(x_{r-1}+0)]$$
$$+ M_{x_r}^{-} f(x) [\phi(x_r) - \phi(x_r - 0)],$$

where $M_x^{+} f(x)$, $M_x^{-} f(x)$ denote the maxima of $f(x)$ at x, on the right and left, respectively, and $U_{x_{r-1}+0}^{x_r-0} f(x)$ is the upper boundary of $f(x)$ in the open interval (x_{r-1}, x_r). In a similar manner it can be proved that

$$L_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x) \leqq m_{x_{r-1}}^{+} f(x) [\phi(x_{r-1}+0) - \phi(x_{r-1})]$$
$$+ L_{x_{r-1}+0}^{x_r-0} f(x) [\phi(x_r - 0) - \phi(x_{r-1}+0)]$$
$$+ m_{x_r}^{-} f(x) [\phi(x_r) - \phi(x_r - 0)],$$

where $m_{x_r}^{+} f(x)$, $m_{x_r}^{-} f(x)$ denote the minima of $f(x)$ at x, on the right and left, respectively, and $L_{x_{r-1}+0}^{x_r-0} f(x)$ is the lower boundary of $f(x)$ in the open interval (x_{r-1}, x_r).

---

* Another proof of this was given by Pollard, *loc. cit.* p. 137.

We now have $\overline{S}_D \geqq \overline{S}_D' \geqq \overline{I}$; $\underline{S}_D \leqq \underline{S}_D' \leqq \underline{I}$, where

$$\overline{S}_D' \equiv \sum_{r=1}^{r=m} U_{x_{r-1}+0}^{x_r-0} f(x) \,.\, \Delta_{x_{r-1}+0}^{x_r-0} \phi(x)$$
$$+ \sum_{r=0}^{r=m-1} M_{x_r}^{+} f(x) \left[\phi(x_r+0) - \phi(x_r)\right]$$
$$+ \sum_{r=1}^{r=m} M_{x_r}^{-} f(x) \left[\phi(x_r) - \phi(x_r-0)\right],$$

and $\underline{S}_D'$ denotes a similar expression, with m written for M, and L for U. It can easily be shewn, by re-introducing ϵ, and proceeding to the limit, that $\overline{S}_D'$ is non-increasing, and $\underline{S}_D'$ non-diminishing as n is increased in D_n. In case $S_{\overline{D}_n}$, $\underline{S}_{\overline{D}_n}$ converge to I, as $\overline{D}_n$ are the successive nets of the system $\overline{D}$, it is clear that $S_{\overline{D}_n}'$, $\underline{S}_{\overline{D}_n}'$ also converge to I. Now $S_{D_n}' - \underline{S}_{D_n}'$ contains the terms

$$\left[M_{x_r}^{+} f(x) - m_{x_r}^{+} f(x)\right]\left[\phi(x_r+0) - \phi(x_r)\right]$$
$$+ \left[M_{x_r}^{-} f(x) - m_{x_r}^{-} f(x)\right]\left[\phi(x_r) - \phi(x_r-0)\right],$$

and this will, for a fixed x_r, be a term of $S_{D_n}' - \underline{S}_{D_n}'$, for all values of n, from and after some fixed integer. Unless this term vanishes it is impossible that $S_{D_n}' - \underline{S}_{D_n}'$ should converge to zero, as $n \sim \infty$. If both $f(x)$ and $\phi(x)$ are discontinuous at x_r, the term can only vanish when, either

$$M_{x_r}^{+} f(x) = m_{x_r}^{+} f(x), \text{ and } \phi(x_r) = \phi(x_r-0),$$

or else

$$M_{x_r}^{-} f(x) = m_{x_r}^{-} f(x), \text{ and } \phi(x_r+0) = \phi(x_r).$$

Thus $\phi(x)$ must be continuous at x_r, on one side, and $f(x)$ must be continuous on the other side.

The following theorem has now been established:

If $(G)\int_a^b f(x)\,\phi(x)\,dx$ *exists, where* $\phi(x)$ *is monotone, all the points at which* $f(x)$ *and* $\phi(x)$ *are both discontinuous must belong to a finite, or enumerable, set of points* H*; and at each point of* H, $f(x)$ *must be continuous on one side and* $\phi(x)$ *must be continuous on the other side. All the points of* H *must be end-points of meshes of a system of nets for which*

$$\sum_{r=1}^{r=m_n} F_{x_{r-1}}^{x_r} f(x) \,.\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

converges to zero, as $n \sim \infty$.

It has also been shewn above that:

In order that $(G)\int_a^b f(x)\,\phi(x)\,dx$ *may exist, it is necessary that a system of nets* $\overline{D}$ *exists, such that*

$$\sum_{r=1}^{r=m_n} F_{x_{r-1}+0}^{x_r-0} f(x) \,.\, \Delta_{x_{r-1}+0}^{x_r-0} \phi(x)$$

converges to zero as $n \sim \infty$, *in the successive nets* $\overline{D}_n$, *of the system* $\overline{D}$.

This is an extension of the result at the end of § 331, for the R-integral.

In case $\int_a^b f(x)\,\phi(x)\,dx$ exists as an RS-integral, $\bar{D}$ may be taken to be any system of nets. Thus we have the theorem that:

For an RS-integral

$$\sum_{r=1}^{r=m_n} F_{x_{r-1}+0}^{x_r-0} f(x) \,.\, \Delta_{x_{r-1}+0}^{x_r-0} \phi(x)$$

converges to zero, for any system of nets.

The generalized RS-integral of a bounded function $f(x)$ with respect to a function of bounded variation $\phi(x)$, or $P(x) - N(x)$ (see § 244) may be defined by

$$(G)\int_a^b f(x)\,d\phi(x) = (G)\int_a^b f(x)\,dP(x) - (G)\int_a^b f(x)\,dN(x),$$

whenever the integrals on the right-hand side exist.

377^1. If, for the net D, or $(a, x_1, x_2, \ldots, b)$, we substitute the net $(a, a+\epsilon, x_1-\epsilon, x_1+\epsilon, x_2-\epsilon, x_2+\epsilon, \ldots, b-\epsilon, b)$, where ϵ is smaller than the least of the numbers $\frac{1}{2}(x_r - x_{r-1})$, we obtain, when ϵ converges to zero, the sums

$$\bar{S}_D'' = \sum_{r=1}^{r=m} U_{x_{r-1}+0}^{x_r-0} f(x) \,.\, \Delta_{x_{r-1}+0}^{x_r-0} \phi(x) + \sum_{r=0}^{r=m} M_{x_r} f(x)\,[\phi(x_r+0) - \phi(x_r-0)],$$

$$\underline{S}_D'' = \sum_{r=1}^{r=m} L_{x_{r-1}+0}^{x_r-0} f(x)\ \Delta_{x_{r-1}+0}^{x_r-0} \phi(x) + \sum_{r=0}^{r=m} m_{x_r} f(x)\,[\phi(x_r+0) - \phi(x_r-0)],$$

where $M_{x_r} f(x)$, $m_{x_r} f(x)$ denote the maxima and minima of $f(x)$ at x_r, and $\phi(x_0 - 0)$ denotes $\phi(a)$, and $\phi(x_m + 0)$ denotes $\phi(b)$. These two sums may be employed to define an integral of $f(x)$ with regard to the monotone function $\phi(x)$ in a manner precisely similar to that employed in the case of the generalized RS-integral. It will be shewn that such a definition leads only to the ordinary RS-integrals, and consequently affords an alternative definition of these integrals. It can easily be shewn, by re-introducing ϵ, and passing to the limit $\epsilon = 0$, that $\bar{S}_D''$ is not increased, and $\underline{S}_D''$ is not diminished, when further points of sub-division of (a, b) are introduced.

If $S_{D_n}'' - \underline{S}_{D_n}''$ converges to zero, for a particular system of nets, it can be shewn, precisely as in § 377, that any point at which $f(x)$, $\phi(x)$ are both discontinuous must be an end-point of meshes of nets of the system.

Also, since $S_D'' - \underline{S}_D''$ contains a term

$$[M_{x_r} f(x) - m_{x_r} f(x)]\,[\phi(x_r+0) - \phi(x_r-0)]$$

which belongs to $\bar{S}_{D_n}'' - \underline{S}_{D_n}''$ for all values of n, from and after some fixed one, this term must be zero. Therefore either $\phi(x)$ or $f(x)$ must be continuous at x_r; and thus $\phi(x)$ and $f(x)$ have no points of discontinuity in common. It is accordingly necessary and sufficient for the existence of the integral, that there be no points in common of discontinuity of $f(x)$ and $\phi(x)$, and that a system of nets exists such that

$$\sum_{r=1}^{r=m} F_{x_{r-1}+0}^{x_r-0} f(x) . \Delta_{x_{r-1}+0}^{x_r-0} \phi(x)$$

should converge to zero for the system of nets.

For a particular net D of the system, we have

$$\sum_{r=1}^{r=m} F_{x_{r-1}+0}^{x_r-0} f(x) . \Delta_{x_{r-1}+0}^{x_r-0} \phi(x) < \tfrac{1}{2}\eta;$$

let another net D', all of whose meshes are $< d$, be superimposed on this net; the inequality will then hold for the new net (D, D'). The number of meshes of D' which are not interior to meshes of D is at most $2m$. At an end-point of a mesh of D,

$$F_{x_r-\epsilon}^{x_r+\epsilon} f(x) . \Delta_{x_r-\epsilon}^{x_r+\epsilon} \phi(x)$$

converges to zero with ϵ, since $f(x)$, $\phi(x)$ are not both discontinuous at x_r. It is accordingly possible so to choose d that the sum of the terms in the expression $\overline{S}_{D'} - \underline{S}_{D'}$, for the net D', which correspond to meshes of D' that are not interior to meshes of D, and of which the number is at most $2m$, is less than $\tfrac{1}{2}\eta$. The sum of the terms of $S_D - \underline{S}_{D'}$, for the remaining meshes of D' is less than $\tfrac{1}{2}\eta$; therefore the sum for all the meshes of D' is less than η, provided d is sufficiently small. Thus, if the integral exists, it is such that

$$\sum_{r=1}^{r=m} F_{x_{r-1}}^{x_r} f(x) . \Delta_{x_{r-1}}^{x_r} \phi(x)$$

for a net such that every mesh is of length $< d$, is $< \eta$. This is the condition for the existence of the RS-integral of $f(x)$ with respect to $\phi(x)$. Therefore

An RS-integral is such that a net exists for which* $S_D'' - \underline{S}_D''$ *is arbitrarily small, when D is properly chosen.*

It has also been shewn that:

The necessary and sufficient conditions for the existence of

$$\int_a^b f(x)\, d\phi(x),$$

as an RS-integral are (1), *that* $f(x)$, $\phi(x)$ *have no points of discontinuity in common, and* (2), *that, if* ϵ *be arbitrarily chosen, a net D exists such that*

$$\sum_{r=1}^{r=m} F_{x_{r-1}+0}^{x_r-0} f(x) . \Delta_{x_{r-1}+0}^{x_r-0} \phi(x) < \epsilon.$$

These conditions are equivalent to the necessary and sufficient condition given in the theorem of § 376¹.

---

* That this is the case was stated to the author by Prof. D. C. Gillespie of Cornell University.

378. It was pointed out in § 252 that the upper and lower variations of a monotone non-diminishing function $\phi(x)$ over any set of points in (a, b) are the exterior and interior measures of the corresponding set of points $\xi\ (= \phi(x))$ in the linear integral of ξ, provided the closed interval $\{\phi(x_1 - 0), \phi(x_1 + 0)\}$ be taken to correspond, on the ξ-segment, to a point x_1, on the x-segment, at which $\phi(x)$ is discontinuous. If $\chi(\xi)$ be defined on the ξ-segment as having the value of $f(x)$ at the point x which corresponds to ξ, the function $\chi(\xi)$ is discontinuous at the point ξ corresponding to a point x at which $f(x)$ is discontinuous and $\phi(x)$ is continuous; but if $\phi(x)$ is discontinuous, $\chi(\xi)$ is constant in the closed interval corresponding to x, and is discontinuous at one or both of the end-points $\phi(x - 0)$, $\phi(x + 0)$ of that interval, in case $f(x)$ is also discontinuous at x.

It will be shewn that:

In order that $(G)\int_a^b f(x)\,d\phi(x)$ *may exist, where* $\phi(x)$ *is monotone, it is necessary that* $\int_\alpha^\beta \chi(\xi)\,d\xi$ *should exist as an R-integral; where* $\alpha = \phi(a)$, $\beta = \phi(b)$. *A fortiori the condition is necessary in order that the RS-integral* $\int_a^b f(x)\,d\phi(x)$ *should exist. When the generalized RS-integral exists, its value is that of the R-integral* $\int_\alpha^\beta \chi(\xi)\,d\xi$.

It has been shewn that the necessary and sufficient condition for the existence of the generalized *RS*-integral is that, when ϵ is arbitrarily prescribed, a net should exist such that

$$\sum_{r=1}^{r=m} F_{x_{r-1}}^{x_r} f(x) \,.\, \Delta_{x_{r-1}}^{x_r} \phi(x) < \epsilon.$$

There corresponds to this net $(a, x_1, x_2, \dots, b)$, a net $(\alpha, \xi_1, \xi_2, \dots, \beta)$ fitted on to the ξ-segment, where $\xi_r = \phi(x_r)$, for $r = 0, 1, 2, \dots, m$. We have then

$$\sum_{r=1}^{r=m} F_{\xi_{r-1}}^{\xi_r} \chi(\xi)\,(\xi_r - \xi_{r-1}) < \epsilon;$$

and that there should exist such a net, for each value of ϵ, is the sufficient condition that the *R*-integral $\int_\alpha^\beta \chi(\xi)\,d\xi$ should exist.

For each value of ϵ, $\int_a^b f(x)\,d\phi(x)$ lies between

$$\sum_{r=1}^{r=m} U_{x_{r-1}}^{x_r} f(x) \,.\, \Delta_{x_{r-1}}^{x_r} \phi(x) \text{ and } \sum_{r=1}^{r=m} L_{x_{r-1}}^{x_r} f(x) \,.\, \Delta_{x_{r-1}}^{x_r} \phi(x)$$

or between

$$\sum_{r=1}^{r=m} U_{\xi_{r-1}}^{\xi_r} \chi(x)\,(\xi_r - \xi_{r-1}) \text{ and } \sum_{r=1}^{r=m} L_{\xi_{r-1}}^{\xi_r} \chi(x)\,(\xi_r - \xi_{r-1});$$

taking a sequence of values of ϵ, converging to zero, and the nets corresponding to these values of ϵ, it is clear that, since the last two of these sums converge, as $\epsilon \sim 0$, to $\int_\alpha^\beta \chi(\xi)\, d\xi$, the two first of these sums converge to the same value, and this is consequently the value of

$$(G) \int_a^b f(x)\, d\phi(x).$$

The converse of the above theorem does not hold good. It is possible that $\int_\alpha^\beta \chi(\xi)\, d\xi$ may exist as an R-integral in cases in which $f(x)$ and $\phi(x)$ have points of discontinuity in common and are such that the condition is not satisfied that, at all such points, $f(x)$ is continuous on one side of the point and $\phi(x)$ on the other side. Thus the R-integral may exist when $(G)\int_a^b f(x)\, d\phi(x)$ does not exist.

For simplicity let us consider the case when $f(x)$, $\phi(x)$ are both discontinuous at a single point $\bar{x}$ in (a, b), and are elsewhere both continuous. The function $\chi(\xi)$ has the constant value $f(\bar{x})$ in the interval

$$[\phi(\bar{x}-0), \phi(\bar{x}+0)];$$

its only discontinuities are at the end-points of this interval, and thus $\int_\alpha^\beta \chi(\xi)\, d\xi$ exists, as an R-integral. A net may be fitted on to the interval $(\phi(a), \phi(b))$ of ξ, such that

$$\sum_{r=1}^{r=m} F_{\xi_{r-1}}^{\xi_r} \chi(\xi)(\xi_r - \xi_{r-1}) < \epsilon.$$

For those values of r which are such that ξ_{r-1}, ξ_r are both in the interval $[\phi(\bar{x}-0), \phi(\bar{x}+0)]$ we have $F_{\xi_{r-1}}^{\xi_r} \chi(\xi) = 0$. Also we may suppose that for some value of r, ξ_r has the value $\phi(\bar{x}-0)$, and for some other value of r, $\xi_r = \phi(\bar{x}+0)$; for this may be secured by dividing those meshes of the net which contain the two points $\phi(\bar{x}-0)$, $\phi(\bar{x}+0)$ into two meshes, and when this is done, the inequality above still holds good. Corresponding to the two meshes $(\xi_s, \phi(\bar{x}-0))$, $(\phi(\bar{x}+0), \xi_t)$, of the net, we have the terms which are equivalent to

$$F_{x_s}^{\bar{x}} f(x)\, [\phi(\bar{x}-0) - \phi(x_s)], \quad F_{\bar{x}}^{x_t} f(x)\, [\phi(x_t) - \phi(\bar{x}+0)],$$

where $\xi_s = \phi(x_s)$, $\xi_t = \phi(x_t)$. These two terms converge to zero as x_s, x_t converge to $\bar{x}$; but the terms corresponding to the meshes $(x_s, \bar{x})$, $(\bar{x}, x_t)$ in the expression for $\overline{S}_D - \underline{S}_D$ in the corresponding net on the x-segment are

$$F_{x_s}^{\bar{x}} f(x)\, [\phi(\bar{x}) - \phi(x_s)], \quad F_{\bar{x}}^{x_t} f(x)\, [\phi(x_t) - \phi(\bar{x})],$$

which do not in general converge to zero as x_s, x_t converge to $\bar{x}$.

In case $\phi(x)$ is of bounded variation and $(G)\int_a^b f(x)\,d\phi(x)$ exists, it is expressible as $(G)\int_a^b f(x)\,dP(x) - (G)\int_a^b f(x)\,dN(x)$, where $P(x)$, $N(x)$ are defined as in § 244. The integral $(G)\int_a^b f(x)\,d\phi(x)$ is then expressible as the difference of the two R-integrals corresponding to $\int_a^b f(x)\,dP(x)$, $\int_a^b f(x)\,dN(x)$, respectively.

It is possible to define the integral $\int_a^b f(x)\,d\phi(x)$, where $\phi(x)$ is monotone, as given by the R-integral $\int_{\phi(a)}^{\phi(b)} \chi(x)\,dx$, whenever the latter exists. This definition is more general than, but includes, the definition given in § 377, of the generalized RS-integral. It may be adopted as the most general definition of an RS-integral, and it will be extended in § 445 to the case of Lebesgue-Stieltjes integrals.

378^1. Let $\phi(x)$ be the R-integral $\int_a^x \phi_1(x)\,dx$, where $\phi_1(x) \geq 0$, in (a, b), and is bounded; the function $\phi(x)$ is then monotone non-diminishing, and is also absolutely continuous (see § 343); thus $\int_a^b f(x)\,d\phi(x)$ exists as a Stieltjes integral, provided $\int_a^\beta \chi(\xi)\,d\xi$ exists, and the two have the same value. In accordance with a theorem proved in § 360, if either of the integrals $\int_a^\beta \chi(\xi)\,d\xi$, $\int_a^b f(x)\,D\phi(x)\,dx$ exists as an R-integral, the other also exists and the two have the same value, since the condition is satisfied (see § 345) that $\int_a^b D\phi(x)$ exists as an R-integral; $D\phi(x)$ denotes any one of the four derivatives of $\phi(x)$ and is equal almost everywhere to $\phi_1(x)$. If either of the integrals

$$\int_a^b f(x)\,D\phi(x)\,dx, \quad \int_a^b f(x)\,\phi_1(x)\,dx$$

exists, then both exist, and they have one and the same value, since $\int_a^b f(x)\,[D\phi(x) - \phi_1(x)]\,dx$ has its integrand bounded, and almost everywhere zero (see § 341).

If $\int_a^b f(x)\,d\phi(x)$ exists, either as an RS-integral or as a generalized RS-integral, then $\int_a^\beta \chi(\xi)\,d\xi$ exists as an R-integral. It then follows that

$\int_a^b f(x)\, D\phi(x)\, dx$ and $\int_a^b f(x)\, \phi_1(x)\, dx$ both exist, and have the same value as $\int_a^b f(x)\, d\phi(x)$.

If it be assumed that $\int_a^b f(x)\, \phi_1(x)\, dx$ exists, then $\int_a^b f(x)\, D\phi(x)\, dx$ exists, and consequently $\int_\alpha^\beta \chi(\xi)\, d\xi$ exists, and has the same value, which is then that of $(G) \int_a^b f(x)\, d\phi(x)$.

The following theorem has now been established:

If $\phi_1(x)$ (≥ 0) *is integrable* (R) *in* (a, b), *and if* $f(x)$ *is bounded in* (a, b), *we have*

$$\int_a^b f(x)\, \phi_1(x)\, dx = (G) \int_a^b f(x)\, d\phi(x), \text{ where } \phi(x) = \int_a^x \phi_1(x)\, dx,$$

provided it is known that either of the integrals exists.

This theorem was stated* by J. M. Whittaker, and he has given a direct proof of it.

379. The fact that the generalized RS-integral of a bounded function $f(x)$ with regard to a bounded monotone function $g(x)$ is represented by an R-integral $\int_\alpha^\beta \chi(\xi)\, d\xi$ may be employed to extend the properties of an R-integral to the case of a generalized RS-integral, or in particular to the case of an RS-integral. The extensions may be made also to the case when $\phi(x)$ is a function of bounded variation, since the generalized RS-integral is then representable as the difference of the R-integrals.

We have thus the theorem.

If $(G) \int_a^b f(x)\, d\phi(x)$ exists, then $(G) \int_a^c f(x)\, d\phi(x)$, $(G) \int_c^b f(x)\, d\phi(x)$ both exist, and

$$(G) \int_a^b f(x)\, d\phi(x) = (G) \int_a^c f(x)\, d\phi(x) + (G) \int_c^b f(x)\, d\phi(x),$$

where $a < c < b$, and $\phi(x)$ is either monotone, or of bounded variation: the converse also holds good. This theorem holds in particular for the Stieltjes integral, provided all three integrals are of this type. It should however be observed that $\int_a^b f(x)\, d\phi(x)$ is not necessarily an RS-integral, when $\int_a^c f(x)\, d\phi(x)$, $\int_c^b f(x)\, d\phi(x)$ are both RS-integrals. For, taking the case in which $\phi(x)$ is monotone, let $f(x)$ be discontinuous at c on the right and continuous on the left, whilst $\phi(x)$ is continuous at c on the right, and discontinuous on the left. In the interval (a, b), $f(x)$ and $\phi(x)$ have the common point of discontinuity c, whereas in the intervals (a, c), (c, b), the point c is not a common point of discontinuity of $f(x)$ and $\phi(x)$; thus the integral over (a, b) cannot be an RS-integral, whilst the integrals over (a, c), (c, b) may both be RS-integrals.

---

* *Proc. Lond. Math. Soc.* (2), vol. xxv (1926), p. 213.

Since $(G)\int_a^x f(x)\,d\phi(x) = \int_a^\xi \chi(\xi)\,d\xi$, it follows that, at a point of continuity of $\phi(x)$ or ξ, $(G)\int_a^x f(x)\,d\phi(x)$ is a continuous function of x. Also, at any point at which $f(x)$ and $\phi(x)$ are continuous,

$$\lim_{h\sim 0} (G)\int_x^{x+h} f(x)\,d\phi(x)\Big/\{\phi(x+h)-\phi(x)\} = f(x);$$

and this is a generalization of the theorem to which it reduces when $\phi(x) = x$, that $\int_a^x f(x)\,dx$ has a differential coefficient equal to $f(x)$ at a point at which $f(x)$ is continuous.

The following theorem will be established:

If $f_1(x)$, $f_2(x)$ both have RS-integrals with respect to the monotone function $\phi(x)$, then $f_1(x)+f_2(x)$ has also an RS-integral with respect to $\phi(x)$, and

$$\int_a^b \{f_1(x)+f_2(x)\}\,d\phi(x) = \int_a^b f_1(x)\,d\phi(x) + \int_a^b f_2(x)\,d\phi(x).$$

Since the variation of $\phi(x)$ is zero over the sets of discontinuities of $f_1(x)$ and of $f_2(x)$, the variation is zero over the set of discontinuities of $f_1(x)+f_2(x)$, and therefore the RS-integral on the left-hand side exists. The equality then follows from the property

$$\int_\alpha^\beta \{\chi_1(\xi)+\chi_2(\xi)\}\,d\xi = \int_\alpha^\beta \chi_1(\xi)\,d\xi + \int_\alpha^\beta \chi_2(\xi)\,d\xi.$$

In general, the properties of the R-integral, given in § 337, may be extended, in the manner indicated above, to the RS-integral, or its generalization.

380. It will be shewn that:

If $\phi(x)$ is continuous and monotone in (a, b), and $f(x)$ is bounded in (a, b), then

$$(G)\overline{\int_a^b} f(x)\,d\phi(x) - (G)\underline{\int_a^b} f(x)\,d\phi(x) = (G)\overline{\int_a^b} \omega[f(x)]\,d\phi(x),$$

where $\omega[f(x)]$ denotes the saltus of $f(x)$ at the point x.

Corresponding to a net fitted on to (a, b) in which the meshes are all of length $< d$, there is a net fitted on to (α, β) of which the lengths are all less than a number d'. The numbers d, d' converge together to zero, on account of the uniform continuity of $\phi(x)$ in (a, b). To a system of nets fitted on to (a, b), there corresponds a system of nets fitted on to (α, β); and conversely. It follows easily that

$$(G)\overline{\int_a^b} f(x)\,d\phi(x) = \overline{\int_\alpha^\beta} \chi(\xi)\,d\xi, \text{ and } (G)\underline{\int_a^b} f(x)\,d\phi(x) = \underline{\int_\alpha^\beta} \chi(\xi)\,d\xi.$$

For a particular system of nets can be fitted on to (a, b), such that, for the system the sums $\overline{S}_{D_n}$, $\underline{S}_{D_n}$ converge, as $n \sim \infty$, to the upper and lower generalized integrals of $f(x)$ with respect to $\phi(x)$. For the corresponding system of nets fitted on to (α, β), the sums

$$\Sigma\Delta_{\xi_{r-1}}^{\xi_r}\chi(\xi)\cdot(\xi_r - \xi_{r-1}), \quad \Sigma\Delta_{\xi_{r-1}}^{\xi_r} f(\xi)\cdot(\xi_r - \xi_{r-1})$$

converge to the upper and lower generalized integrals of $f(x)$ with regard to ξ. But, on account of the property of the upper and lower R-integrals given in § 331, these sums converge to

$$\overline{\int_\alpha^\beta} \chi(\xi)\,d\xi, \quad \underline{\int_\alpha^\beta} \chi(\xi)\,d\xi,$$

respectively. Therefore

$$(G)\overline{\int_a^b} f(x)\,d\phi(x) = \overline{\int_\alpha^\beta} \chi(\xi)\,d\xi, \text{ and } (G)\underline{\int_a^b} f(x)\,d\phi(x) = \underline{\int_\alpha^\beta} \chi(\xi)\,d\xi.$$

It has been shewn in § 334, that

$$\overline{\int_\alpha^\beta} \chi(\xi)\,d\xi - \underline{\int_\alpha^\beta} \chi(\xi)\,d\xi = \overline{\int_\alpha^\beta} \omega[\chi(\xi)]\,d\xi,$$

where $\omega[\chi(\xi)]$ is the saltus of $\chi(\xi)$ at ξ; and this is equal to the saltus of $f(x)$ at x. Moreover, applying the above result to the function $\omega[f(x)]$, we have

$$\overline{\int_a^b} \omega[f(x)]\,d\phi(x) = \overline{\int_\alpha^\beta} \omega[\chi(\xi)]\,d\xi.$$

The truth of the theorem now follows from the theorem of § 334.

This theorem has been established* otherwise by Pollard.

If $\chi_1(\xi)$ be integrable (R) in the interval (α_1, β_1), and is zero outside that interval; and if $\chi_2(\xi)$ is integrable (R) in the interval (α_2, β_2), and is zero outside that interval, we find by applying the theorem given in § 351 to an interval (A, B) which contains both the intervals (α_1, β_1), (α_2, β_2), that

$$\int_{\alpha_1}^{\beta_1} \chi_1(\xi)\left[\int_{\alpha_2}^{\xi} \chi_2(\xi)\,d\xi\right] d\xi + \int_{\alpha_2}^{\beta_2} \chi_2(\xi)\left[\int_{\alpha_1}^{\xi} \chi_1(\xi)\,d\xi\right] d\xi$$
$$= \int_{\alpha_1}^{\beta_1} \chi_1(\xi)\,d\xi \int_{\alpha_2}^{\beta_2} \chi_2(\xi)\,d\xi.$$

---

* *Quarterly Journal*, vol. XLIX (1923), p. 130.

Let $\phi(x)$, $\psi(x)$ be monotone bounded functions defined in the interval (a, b); and let $f(x)$, $g(x)$ be bounded functions defined in the same interval. Let $f(x) \equiv \chi_1(\xi)$, where $\xi = \phi(x)$, and let $g(x) \equiv \chi_2(\xi)$, where $\xi = \psi(x)$; also let $\alpha_1 = \phi(a)$, $\beta_1 = \phi(b)$, $\alpha_2 = \psi(a)$, $\beta_2 = \psi(b)$; the above formula is then equivalent to

$$\int_a^b f(x)\left[\int_a^x g(x)\,d\psi(x)\right]d\phi(x) + \int_a^b g(x)\left[\int_a^x f(x)\,d\phi(x)\right]d\psi(x)$$
$$= \int_a^b f(x)\,d\phi(x)\int_a^b g(x)\,d\psi(x) \text{ ....(A)},$$

provided all the integrals in this expression exist as generalized *RS*-integrals. This formula (A) corresponds, for generalized *RS*-integrals, to the formula of § 351, for *R*-integrals. It clearly holds when $\phi(x)$, $\psi(x)$ are functions of bounded variation, as the result is then obtained by applying (A) to the four pairs of monotone functions

$$\phi_1(x), \psi_1(x);\ \phi_2(x), \psi_2(x);\ \phi_1(x), \psi_2(x);\ \phi_2(x), \psi_1(x),$$

where $\phi(x) = \phi_1(x) - \phi_2(x)$, $\psi(x) = \psi_1(x) - \psi_2(x)$.

Let us now suppose that $\int_a^b f(x)\,d\phi(x)$, $\int_a^b g(x)\,d\psi(x)$ exist as *RS*-integrals. In case $\phi(x)$, $\psi(x)$ have no points of discontinuity in common,

$$\int_a^b f(x)\left[\int_a^x g(x)\,d\psi(x)\right]d\phi(x)$$

exists as an *RS*-integral. For, in order that this may be so, it is only necessary that the measure of $\phi(x)$ over the set of discontinuities of the integrand should be zero. Since $\int_a^x g(x)\,d\psi(x)$ is discontinuous only at the points of discontinuity of $\psi(x)$, which form an enumerable set, the measure of $\phi(x)$ over this set is zero (see § 252); also the measure of $\phi(x)$ over the set of points of discontinuity of $f(x)$ is zero, since $f(x)$ has an *RS*-integral with respect to $\phi(x)$. It follows that the measure of $\phi(x)$ over the set of points of discontinuity of

$$f(x)\left[\int_a^x g(x)\,d\psi(x)\right]$$

is zero, and therefore

$$\int_a^b f(x)\left[\int_a^x g(x)\,d\psi(x)\right]d\phi(x)$$

is an *RS*-integral. In the same way it is seen that the second integral in (A) is an *RS*-integral.

If $\phi(x)$, $\psi(x)$ had a common point of discontinuity, $\phi(x)$ and

$$f(x)\int_a^x g(x)\,d\psi(x)$$

would have a point of discontinuity in common, and thus the integrals in (A) would not exist as *RS*-integrals.

It has now been shewn that:

If $\phi(x)$, $\psi(x)$ are functions of bounded variation in (a, b), and have no point of discontinuity in common; and if $f(x)$ is bounded and possesses an RS-integral with respect to $\phi(x)$; and if $g(x)$ is bounded and possesses an RS-integral with respect to $\psi(x)$; then the formula (A) *holds good.*

This theorem was given* by Hardy who furnished a direct proof of it.

EXAMPLES

1. Let $\phi(x) = 0$, for $0 \leqq x < \frac{1}{2}$; $\phi(x) = 1$, for $\frac{1}{2} \leqq x \leqq 1$; and $f(x) = 1$, for $x \neq \frac{1}{2}$; $f(\frac{1}{2}) = 0$. The function $f(x)$ being discontinuous on both sides at the point $\frac{1}{2}$, where $\phi(x)$ is also discontinuous, $(G)\int_0^1 f(x)\,d\phi(x)$ does not exist, but $\int_0^1 \chi(\xi)\,d\xi$ exists, and has the value zero; the interval $(0, 1)$ being $\phi(\frac{1}{2} - 0)$. $\phi(\frac{1}{2} + 0)$, in which $\chi(\xi) = 0$.

2. Let $\phi(x) = x$, for $0 \leqq x < 1$; $\phi(x) = 2x$, for $1 \leqq x \leqq 2$. $f(x) = 1$, for $0 \leqq x \leqq 1$; $f(x) = 2$, for $1 < x \leqq 2$. The functions $f(x)$, $\phi(x)$ are both discontinuous at $x = 1$, but $f(x)$ is continuous on the left, and $\phi(x)$ is continuous on the right of the point. The generalized *RS*-integral exists. Moreover, in the intervals $(0, 1)$, $(1, 2)$, $f(x)$ and $\phi(x)$ have no point of discontinuity in common In these intervals the *RS*-integrals exist, although the integral over the interval $(0, 2)$ is not an *RS*-integral.

THE RIEMANN-STIELTJES INTEGRAL FOR FUNCTIONS OF TWO VARIABLES

381. Let a bounded function $f(x^{(1)}, x^{(2)})$, defined in a cell $(a^{(1)}, a^{(2)}: b^{(1)}, b^{(2)})$, be bounded in that cell, and let $\phi(x^{(1)}, x^{(2)})$ be a function defined in the same cell, and quasi-monotone, in accordance with the definition in § 255; thus, employing the notation there introduced, $\Delta_{(x^{(1)}, x^{(2)})}^{(\bar{x}^{(1)}, \bar{x}^{(2)})} \phi(x^{(1)}, x^{(2)}) \geqq 0$, provided $\bar{x}^{(1)} \geqq x^{(1)}$, $\bar{x}^{(2)} \geqq x^{(2)}$.

Let a net be fitted on to the cell in which the functions are defined, and let $\delta_1, \delta_2, \ldots \delta_m$ denote the meshes of the net, and $\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)})$ denote $\Delta_{(\alpha_r^{(1)}, \alpha_r^{(2)})}^{(\beta_r^{(1)}, \beta_r^{(2)})} \phi(x^{(1)}, x^{(2)})$, where $(\alpha_r^{(1)}, \alpha_r^{(2)}; \beta_r^{(1)}, \beta_r^{(2)})$ is the mesh δ_r; moreover, let $U(\delta_r)$, $L(\delta_r)$ denote the upper, and the lower, boundary of $f(x^{(1)}, x^{(2)})$ in the cell δ_r.

The sum $\sum\limits_{r=1}^{r=m} U(\delta_r)\,\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)})$ has a lower boundary, when all possible nets are taken into account; and this lower boundary is said to define

$$\overline{\int}_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\,d\phi(x^{(1)}, x^{(2)}),$$

the upper integral of $f(x^{(1)}, x^{(2)})$ with respect to $\phi(x^{(1)}, x^{(2)})$, in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

---

* *Messenger of Math.* vol. XLVIII (1918), p. 90.

Similarly, the *lower integral of* $f(x^{(1)}, x^{(2)})$ *with respect to* $\phi(x^{(1)}, x^{(2)})$ is defined as the upper boundary of $\sum_{r=1}^{r=m} L(\delta_r)\,\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)})$, for all possible nets fitted on to the fundamental cell, and this lower integral is denoted by

$$\underline{\int}_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d\phi(x^{(1)}, x^{(2)}).$$

In case the upper and lower integrals of $f(x^{(1)}, x^{(2)})$ *with respect to* $\phi(x^{(1)}, x^{(2)})$ *are equal, their common value* $\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d\phi(x^{(1)}, x^{(2)})$ *is taken to define the Riemann-Stieltjes, or RS-integral, of* $f(x^{(1)}, x^{(2)})$ *with respect to* $\phi(x^{(1)}, x^{(2)})$.

In case $\phi(x^{(1)}, x^{(2)})$ is simply the product $x^{(1)}.x^{(2)}$, we see that $\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)})$ reduces to $(\beta_r^{(1)} - \alpha_r^{(1)})(\beta_r^{(2)} - \alpha_r^{(2)})$, the measure of the cell δ_r; the *RS*-integral is then the *R*-integral, as defined in § 338.

If $\phi(x^{(1)}, x^{(2)})$ be a function of bounded variation, in accordance with the definition of Hardy and Krause, given in § 254, it is expressible as the difference of two quasi-monotone functions $\overline{P}(x^{(1)}, x^{(2)})$, $\overline{N}(x^{(1)}, x^{(2)})$, (see §§ 254, 255); which are also monotone. The *RS*-integral of $f(x^{(1)}, x^{(2)})$, with respect to $\phi(x^{(1)}, x^{(2)})$, is defined by

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d\phi(x^{(1)}, x^{(2)})$$
$$= \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d\overline{P}(x^{(1)}, x^{(2)}) - \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, d\overline{N}(x^{(1)}, x^{(2)}),$$

whenever the two latter integrals exist.

It can be shewn that a continuous function $f(x^{(1)}, x^{(2)})$ certainly has an *RS*-integral with respect to a function $\phi(x^{(1)}, x^{(2)})$, of bounded variation. We need only consider the case in which $\phi(x^{(1)}, x^{(2)})$ is quasi-monotone. In case the spans of all the meshes of a net, fitted on to the fundamental cell, are sufficiently small,

$$\sum_{r=1}^{r=m} U(\delta_r)\,\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)}) \text{ exceeds } \sum_{r=1}^{r=m} L(\delta_r)\,\Delta_{\delta_r}\phi(x^{(1)}, x^{(2)})$$

by less than $\epsilon\Delta_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})}\phi(x^{(1)}, x^{(2)})$, which is an arbitrarily small number. Consequently the upper and lower integrals of $f(x^{(1)}, x^{(2)})$ with respect to $\phi(x^{(1)}, x^{(2)})$ have the same value, and the *RS*-integral of $f(x^{(1)}, x^{(2)})$ therefore exists.

If $f(x^{(1)}, x^{(2)})$ have discontinuities, the criterion for the existence of the integral with respect to the function $\phi(x^{(1)}, x^{(2)})$ is contained in the following theorem*, which may be established by a modification of the method of § 377.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p. 281; also *Proc. Roy. Soc.* vol. XCIII (1917), p. 28.

The necessary and sufficient condition that the bounded function $f(x^{(1)}, x^{(2)})$ *should possess an RS-integral with respect to the function* $\phi(x^{(1)}, x^{(2)})$, *of bounded variation, in accordance with the definition in* § 254, *is that the variation of* $\phi(x^{(1)}, x^{(2)})$ *over the set of points of discontinuity of* $f(x^{(1)}, x^{(2)})$ *should be zero.*

The variation of a quasi-monotone increasing function $\phi(x^{(1)}, x^{(2)})$, over a set of points G, is here taken to denote the lower boundary of $\Sigma \Delta_{(\delta)} \phi(x^{(1)}, x^{(2)})$, where the summation is taken for all cells of a set which contain in them all the points of G, the lower boundary being taken for all such sets. The variation of a function $\phi(x^{(1)}, x^{(2)})$, of bounded variation, over the set G, is taken to be the sum of the variations over G of the two monotone increasing functions, $\overline{P}(x^{(1)}, x^{(2)})$, $\overline{N}(x^{(1)}, x^{(2)})$, the difference of which is $\phi(x^{(1)}, x^{(2)})$ (see § 254).

Further properties of the *RS*-integral for functions of two variables will be given in § 448[1].

CHAPTER VII

THE LEBESGUE INTEGRAL

382. The definition of a definite integral which has been introduced into Analysis by Lebesgue is of much wider scope than the definition of Riemann. The theory of Lebesgue integration has as its foundation the conception of the measure of a set of points, in the sense in which the term is employed by Lebesgue. An account of this theory of measure has been given in Chapter III. In Riemann integration, the domain over which the integral is taken is divided into a finite number of intervals, or of cells, and the integral is defined as the limit of the Riemann sum for this set of intervals, or cells. In Lebesgue integration, on the other hand, the domain over which the integral is taken is divided into a number of measurable sets of points, having a certain property relative to the function to be integrated, and the integral is defined as the limit of a certain sum taken for all these measurable sets of points, as the number of sets is indefinitely increased. The distinction between the Lebesgue integral and the Riemann integral rests essentially upon the difference between the two modes of dividing the domain of integration into sets of points.

MEASURABLE FUNCTIONS

383. The definition of Lebesgue is applicable to functions belonging to the family of measurable functions, in one, or more, dimensions. A measurable function $f(x)$ has already been defined, in § 295, as such that the set of points x, for which $f(x) > A$, is measurable, whatever real number A may be. This definition is applicable, whether x denote a point of a linear set, or denote a point $(x^{(1)}, x^{(2)}, \dots x^{(p)})$, in any number p, of dimensions.

If $f(x)$ be a measurable function, defined at each point of a given domain, the sets of points for which

$$A < f(x) < B;\ A \leq f(x) < B;\ A \leq f(x) \leq B;\ f(x) < A;\ f(x) \leq A$$

are all measurable, whatever real numbers A and B may denote, provided $A < B$.

In the first place, the domain for which $f(x)$ is defined, and for which it has a definite value at each point, is measurable. For let A have the values $-N_1, -N_2, \dots -N_n, \dots$ successively, of a sequence such that N_n increases indefinitely as $n \sim \infty$. The set E_n, for which $f(x) > -N_n$, is measurable, by hypothesis, for every value of n. The domain for which

$f(x)$ is defined is the outer limiting set of the sequence $\{E_n\}$, of measurable sets, and is therefore itself measurable. The set of points for which $f(x) \leqq A$, being complementary, relatively to the domain of the function, to the measurable set for which $f(x) > A$, is measurable. If $\{A_n\}$ be a monotone increasing sequence of numbers converging to A, all the sets for which $f(x) \leqq A_n$ are measurable, and their outer limiting set, for which $f(x) < A$, is consequently measurable. As the sets for which $f(x) < A$ and $f(x) \leqq A$ are measurable, it follows that the set for which $f(x) = A$ is measurable.

The sets for which $f(x) < B$, and $f(x) < A$, being measurable, their difference, the set for which $A \leqq f(x) < B$, is also measurable. The other results in the theorem follow at once.

A function $f(x)$ is measurable if the set of points x is measurable, for which $\alpha < f(x) < \beta$, for every pair α, β, of real numbers which belong to a given set, everywhere dense in the indefinite interval $(-\infty, \infty)$. The given set may be taken to be enumerable.

Let A and B be any pair of real numbers such that $A < B$. The number A can be expressed as the upper limit of a sequence $\{\alpha_n\}$ of increasing numbers, all of which belong to the given everywhere dense set, and the number B can be expressed as the lower limit of a similar sequence $\{\beta_n\}$ of diminishing numbers. The set e_n for which $\alpha_n < f(x) < \beta_n$ is measurable, for each value of n, the inner limiting set $\{e_n\}$, of the sequence, is the set for which $A \leqq f(x) \leqq B$; and this set is consequently measurable. Since this is the case whatever values A and B may have, it is seen, as before, that $f(x)$ is measurable.

A function $f(x)$ is said to be measurable (B), if the set of points for which $f(x) > A$ is measurable (B), whatever value A may have.

The proofs given above shew that the sets for which

$$A < f(x) < B;\ A \leqq f(x) < B;\ A \leqq f(x) \leqq B;\ f(x) < A;\ f(x) \leqq A$$

are all measurable (B).

384. *If $\phi_1, \phi_2, \ldots \phi_n$ be a finite set of functions that are measurable in a measurable domain G, linear, or of higher dimensions, and if $F(\phi_1, \phi_2, \ldots \phi_n)$ be a function that is continuous relatively to $(\phi_1, \phi_2, \ldots \phi_n)$, for all values of $\phi_1, \phi_2, \ldots \phi_n$, then $F(\phi_1, \phi_2, \ldots \phi_n)$ is measurable in the given domain.*

First, let us assume that all the functions $\phi_1, \phi_2, \ldots \phi_n$ are bounded in the given domain for which they are defined; suppose their values all to be in the interval $(-N, N)$. Let a net $(c_0, c_1, \ldots c_m)$ be fitted on to the linear interval $(-N, N)$, where $c_0 = -N$, $c_m = N$, and suppose the breadth $c_r - c_{r-1}$ of each mesh to be less than the positive number η. Let the function ψ_s be defined, corresponding to each function ϕ_s $(s = 1, 2, 3, \ldots n)$

by the conditions $\psi_s = c_{r-1}$ at every point at which $c_{r-1} \leqq \phi_s < c_r$, for $r = 1, 2, 3, \ldots m$, and $\psi_s = c_m$ where $\phi_s = c_m$. We have then

$$0 \leqq \phi_s - \psi_s < \eta,$$

and the function ψ_s taking only the values in the finite set $c_0, c_1, \ldots c_m$, this function is measurable in the given domain.

Since $F(\phi_1, \phi_2, \ldots \phi_n)$ is continuous in the closed domain,

$$(-N, -N, \ldots; N, N, \ldots),$$

we have $$| F(\phi_1, \phi_2, \ldots \phi_n) - F(\psi_1, \psi_2, \ldots \psi_n) | < \epsilon,$$

provided η be taken sufficiently small; the number ϵ being arbitrarily chosen. The function $F(\psi_1, \psi_2, \ldots \psi_n)$ has only a finite set of values, and is measurable If U and L are its upper and lower boundaries, we have

$$L - \epsilon < F(\phi_1, \phi_2, \ldots \phi_n) < U + \epsilon$$

in the whole domain. Let A and B be any two numbers in the interval (L, U), then the set of points for which $A < F(\psi_1, \psi_2, \ldots \psi_n) < B$ is measurable.

Now let ϵ have successively the values in a sequence $\{\epsilon_t\}$ which converges to zero, then there exists a corresponding sequence $\{\eta_t\}$, of values of η, which converges to zero.

The set of points E_t, for which $A < F(\psi_1, \psi_2, \ldots \psi_n) < B$, is measurable, for each value of η_t, in $\{\eta_t\}$. Each point of the set for which

$$A < F(\phi_1, \phi_2, \ldots \phi_n) < B$$

belongs to all the measurable sets E_t, from and after some particular value of t, and therefore, by a theorem established in § 131, the set is measurable. It has thus been shewn that $F(\phi_1, \phi_2, \ldots \phi_n)$ is measurable in the domain for which the functions are defined.

Next, let the functions $\phi_1, \phi_2, \ldots \phi_n$ be unbounded. Let $\phi_r^{(N)}$ be defined by the conditions $\phi_r^{(N)} = \phi_r$, when $N \geqq \phi_r \geqq -N$; $\phi_r^{(N)} = N$, when $\phi_r > N$; and $\phi_r^{(N)} = -N$, when $\phi_r < -N$. From what has been proved above, we see that the function $F(\phi_1^{(N)}, \ldots \phi_n^{(N)})$ is measurable. Let N have successively the values in a divergent sequence $\{N_t\}$ of increasing numbers. Each point of the set for which $A < F(\phi_1, \phi_2, \ldots \phi_n) < B$ belongs to all the measurable sets for which $A < F(\phi_1^{(N_t)}, \phi_2^{(N_t)}, \ldots \phi_n^{(N_t)}) < B$, from and after some particular value of t. It thus follows that the set is measurable. Therefore the theorem holds when $\phi_1, \phi_2, \ldots \phi_n$ are unbounded.

In particular we have the theorem that:

The sum, or the product, of any finite number of measurable functions, defined in a measurable domain of any number of dimensions, is a measurable function.

If all the functions $\phi_1, \phi_2, \ldots \phi_n$ are measurable (B), the function $F(\phi_1, \phi_2, \ldots \phi_n)$ is measurable (B). For the sets employed in the above proof are all measurable (B).

THE LEBESGUE INTEGRAL OF A MEASURABLE FUNCTION

385. Let us consider a measurable set e, bounded or unbounded, but of finite measure, and let the function $f(x) = 1$ at all points of e. The measure $m(e)$, of the set e, is said to define the integral $\int_{(e)} f(x)\,dx$, of the function $f(x)$, over the set e.

If c be any positive, or negative, number, and $f(x) = c$ at all points of e, $\int_{(e)} f(x)\,dx$ is defined to be the number $cm(e)$. The values of $f(x)$ at points that do not belong to e are irrelevant. Next, let $e_1, e_2, \dots e_n$ be a finite number of measurable sets, no two of which have a point in common.

Let $f(x) = c_r$ at the points of e_r, for $r = 1, 2, 3, \dots n$; where $c_1, c_2, \dots c_n$ are assigned numbers. Then, if E denote the set $e_1 + e_2 + \dots + e_n$, the integral $\int_{(E)} f(x)\,dx$, of $f(x)$, over the set E, is defined as the sum

$$\sum_{r=1}^{r=n} c_r m(e_r).$$

Next, let $f(x)$ be a measurable function, defined for the points of a measurable set E, of finite measure, and bounded in that set. Let U, L denote respectively the upper, and the lower, boundary of $f(x)$ in E.

Let the interval (L, U) be divided into parts

$$(a_0, a_1),\ (a_1, a_2),\ \dots\ (a_{n-1}, a_n),$$

where $a_0 = L$, $a_n = U$, and such that the greatest of these parts $a_r - a_{r-1}$, for $r = 1, 2, 3, \dots n$, is $< \eta$.

Let e_r be that measurable part of E, for all points of which

$$a_{r-1} \leq f(x) < a_r,$$

where $r = 1, 2, 3, \dots n-1$; and let e_n be that part of $f(x)$ for the points of which $a_{n-1} \leq f(x) \leq U$.

Let $\phi_\eta(x)$ be the function which has the value a_{r-1}, at all points of e_r, for $r = 1, 2, 3, \dots n$; and let $\psi_\eta(x)$ be the function which has the value a_r, at all points of e_r, for $r = 1, 2, 3, \dots n$.

We have then, in accordance with the above definition,

$$\int_{(E)} \phi_\eta(x)\,dx = \sum_{r=1}^{r=n} a_{r-1} m(e_r),$$

$$\int_{(E)} \psi_\eta(x)\,dx = \sum_{r=1}^{r=n} a_r m(e_r),$$

and therefore $$0 \leq \int_{(E)} \psi_\eta(x)\,dx - \int_{(E)} \phi_\eta(x)\,dx < \eta m(E).$$

If the interval (L, U) be successively sub-divided by introducing further points of division such that the corresponding values of η form a sequence $\{\eta_m\}$ of diminishing numbers for which $\eta_m \sim 0$, as $m \sim \infty$, it is seen that

$$\int_{(E)} \phi_{\eta_m}(x)\, dx$$

does not diminish, and that $\int_{(E)} \psi_{\eta_m}(x)\, dx$ does not increase, as $m \sim \infty$. These two sets of numbers consequently converge to a common limit. This limit $\lim\limits_{m\sim\infty} \int_{(E)} \phi_{\eta_m}(x)\, dx \equiv \lim\limits_{m\sim\infty} \int_{(E)} \psi_{\eta_m}(x)\, dx$ is defined to be the value of the Lebesgue integral $\int_{(E)} f(x)\, dx$, of $f(x)$, taken over the measurable set E.

In order to justify this definition, it is necessary to prove that the number so defined is independent of the particular mode in which the interval (L, U) has been successively sub-divided.

Let $\bar{\phi}_\eta(x)$, $\bar{\psi}_\eta(x)$ be the functions which correspond, in a second mode of sub-division, to $\phi_\eta(x)$, $\psi_\eta(x)$ defined above. It is easily seen that there is no loss of generality in taking the sequence $\{\eta_n\}$ to be the same in the two cases

Suppose the two sub-divisions of (L, U), corresponding to η_m, to be superimposed, and let $\bar{\bar{\phi}}_{\eta_m}(x)$ be the function corresponding to $\phi_{\eta_m}(x)$ and $\bar{\phi}_{\eta_m}(x)$. We have then

$$0 \leq \int_{(E)} \bar{\bar{\phi}}_{\eta_m}(x)\, dx - \int_{(E)} \phi_{\eta_m}(x)\, dx < \eta_m m(E),$$

$$0 \leq \int_{(E)} \bar{\bar{\phi}}_{\eta_m}(x)\, dx - \int_{(E)} \bar{\phi}_{\eta_m}(x)\, dx < \eta_m m(E),$$

and therefore
$$\left| \int_{(E)} \phi_{\eta_m}(x)\, dx - \int_{(E)} \bar{\phi}_{\eta_m}(x)\, dx \right| < \eta_m m(E).$$

As $m \sim \infty$, $\eta_m m(E) \sim 0$; and therefore

$$\lim_{m\sim\infty} \int_{(E)} \phi_{\eta_m}(x)\, dx = \lim_{m\sim\infty} \int_{(E)} \bar{\phi}_{\eta_m}(x)\, dx.$$

Thus the number by which the Lebesgue integral is defined is independent of the mode of sub-division of the set E.

Accordingly, the definition may be stated as follows:

The Lebesgue integral $\int_{(E)} f(x)\, dx$ *of the bounded measurable function* $f(x)$, *taken over the measurable set E which is of finite measure, is defined as the limit to which either* $\sum\limits_{r=1}^{r=n} a_{r-1} m(e_r)$, *or* $\sum\limits_{r=1}^{r=n} a_r m(e_r)$, *tends, as the greatest* *of the numbers* $a_r - a_{r-1}$ *converges to zero: where* (L, U) *is divided into* n *parts* (a_{r-1}, a_r), *for* $r = 1, 2, 3, \ldots n$; *and* $a_0 = L$, $a_n = U$; *and where* e_r *is that set of points of* E, *for all of which*

$$a_{r-1} \leq f(x) < a_r; \quad r = 1, 2, 3, \ldots n - 1;$$

and e_n *is that set for which* $a_{n-1} \leq f(x) \leq a_n$. U *and* L *denote the upper and lower boundaries of* $f(x)$ *in* E.

In case the set E is a set in p-dimensional space, $f(x)$ is used to denote $f(x^{(1)}, x^{(2)}, \ldots x^{(p)})$.

If the set E consists of the points of a finite interval (a, b), or of a cell (a, b), the integral over E is denoted by $\int_a^b f(x)\, dx$,

or
$$\int_{(a^{(1)}, a^{(2)}, \ldots a^{(p)})}^{(b^{(1)}, b^{(2)}, \ldots b^{(p)})} f(x^{(1)}, x^{(2)}, \ldots x^{(p)})\, d(x^{(1)}, x^{(2)}, \ldots x^{(p)}).$$

It will be observed that the function $\phi_\eta(x)$ is a function which takes only a finite number of values, for all values of x in the set E; and that it is such that $0 \leq f(x) - \phi_\eta(x) < \eta$

The integral $\int_{(E)} f(x)\, dx$ is thus defined as the limit of the non-diminishing sequence of numbers

$$\int_{(E)} \phi_{\eta_1}(x)\, dx, \int_{(E)} \phi_{\eta_2}(x)\, dx, \ldots \int_{(E)} \phi_{\eta m}(x)\, dx, \ldots,$$

where $\eta_1, \eta_2, \ldots \eta_m, \ldots$ form a sequence of diminishing numbers that converges to zero, as $m \sim \infty$.

386. Next, let the measurable function $f(x)$ be unbounded in the measurable set E, of finite measure, and such that $f(x) \geq 0$, in E.

Let $f_N(x)$ be defined as a bounded function in E, by the specifications,

$$f_N(x) = f(x), \text{ where } f(x) \leq N,$$
$$f_N(x) = N, \text{ where } f(x) > N.$$

The number N is any assigned positive number.

If there be assigned to N the values in an increasing sequence $N_1, N_2, \ldots$ without upper limit, the integral $\int_{(E)} f_N(x)\, dx$ is a number which does not diminish as N has successively the values in the sequence. It follows that $\int_{(E)} f_N(x)\, dx$ either converges to a definite upper limit, or that it increases indefinitely as N does so.

When $\lim\limits_{N \sim \infty} \int_{(E)} f_N(x)\, dx$ *exists as a definite number, the integral* $\int_{(E)} f(x)\, dx$ *is defined by that number.*

The value of the integral is easily seen to be independent of the particular sequence of values of N. Thus

$$\int_{(E)} f(x)\,dx = \lim_{N\sim\infty} \int_{(E)} f_N(x)\,dx.$$

If $f(x)$ be $\leqq 0$, in E, $\int_{(E)} f(x)\,dx$ is similarly defined as

$$\lim_{N\sim\infty} \int_{(E)} f_{-N}(x)\,dx,$$

where $f_{-N}(x) = f(x)$, for points x at which $f(x) \geqq -N$, and $f_{-N}(x) = -N$, when $f(x) < -N$.

Lastly, let the unbounded function $f(x)$, defined in E, have both positive and negative values.

Let $f(x) = f^+(x) - f^-(x)$, where $f^+(x) = f(x)$, when $f(x) \geqq 0$, and $f^+(x) = 0$, when $f(x) < 0$; with a similar definition for $f^-(x)$. Both the functions $f^+(x)$, $f^-(x)$ are measurable.

The integral $\int_{(E)} f(x)\,dx$ *is defined as the difference,*

$$\int_{(E)} f^+(x)\,dx - \int_{(E)} f^-(x)\,dx.$$

provided both of the latter integrals exist as finite numbers.

When a measurable function $f(x)$, *of one or more variables, is such that* $\int_{(E)} f(x)\,dx$ *exists as a finite number,* $f(x)$ *is said to be summable in the set* E.

A measurable function is always summable if it be bounded, but not necessarily so if it be unbounded.

A function $f(x)$ that is summable in E is also said to be *integrable* (L) in E, and the integral $\int_{(E)} f(x)\,dx$ is termed the *Lebesgue integral*, or shortly the *L-integral* of $f(x)$ in E.

387. It will be observed that, in accordance with the above definition, in order that a measurable function $f(x)$, defined for the measurable set of points E, may be summable in E, it is necessary that each of the two functions $f^+(x)$, $f^-(x)$ should be summable, that is, each one of them must be integrable (L) in E. Thus the two limits

$$\lim_{N\sim\infty} \int_{(E)} f_N^+(x)\,dx, \quad \lim_{N\sim\infty} \int_{(E)} f_N^-(x)\,dx$$

must be both finite.

The definition may accordingly be stated in the following form, which is a generalization of a definition* due to de la Vallée Poussin, originally applicable to the case in which $f_N^+(x)$, $f_N^-(x)$ are restricted to be integrable (R) in a cell or interval (a, b).

---

* *Liouville's Journal* (4), vol. VIII (1892), p. 427.

If the unbounded measurable function $f(x)$ be defined in a measurable set E, of finite measure, and N and N' denote two positive numbers, let $f_{N,N'}(x)$ be such that $f_{N,N'}(x) = f(x)$, at points such that $-N' \leqq f(x) \leqq N$; and $f_{N,N'}(x) = N$ at points such that $f(x) > N$; and $f_{N,N'}(x) = -N'$, when $f(x) < -N'$; then $\int_{(E)} f(x)\,dx$ is defined as the double limit of $\int_{(E)} f_{N,N'}(x)\,dx$, as N and N' diverge, independently of one another, to ∞, whenever this double limit exists as a definite number.

In accordance with this definition, when $f(x)$ is summable, so also is $|f(x)|$; for $|f(x)| = f_N^+(x) + f_N^-(x)$, and therefore $\int_{(E)} |f(x)|\,dx$ is the sum of

$$\lim_{N\sim\infty} \int_{(E)} f_N^+(x)\,dx \text{ and } \lim_{N\sim\infty} \int_{(E)} f_N^-(x)\,dx,$$

each of which is finite.

Since the existence of $\int_{(E)} f(x)\,dx$ involves that of $\int_{(E)} |f(x)|\,dx$, the L-integral $\int_{(E)} f(x)\,dx$ is said to be an absolutely convergent integral, thus:

The L-integral of an unbounded summable function is an absolutely convergent integral.

It is easily seen that $\left|\int_{(E)} f(x)\,dx\right| \leqq \int_{(E)} |f(x)|\,dx$; for the absolute value of the integral on the left-hand side is

$$\leqq \int_{(E)} f^+(x)\,dx + \int_{(E)} f^-(x)\,dx.$$

If $f^{(1)}(x) \geqq f^{(2)}(x) \geqq 0$, the function $f^{(1)}(x)$ being summable in E, then the function $f^{(2)}(x)$, assumed to be measurable in E, is also summable. For we see that $f_N^{(1)}(x)\,dx \geqq f_N^{(2)}(x)\,dx$, hence

$$\int_{(E)} f_N^{(2)}(x)\,dx \leqq \int_{(E)} f_N^{(1)}(x)\,dx,$$

and it then follows that $\int_{(E)} f^{(2)}(x)\,dx$ exists, and is $\leqq \int_{(E)} f^{(1)}(x)\,dx$.

388. The above definition of the L-integral, when $f(x) \geqq 0$, in E, and the function $f(x)$ is unbounded, may be replaced by another definition which we proceed to obtain.

Let $a_0, a_1, a_2, \ldots a_n, \ldots$ be an increasing sequence of numbers, such that a_n has no upper boundary, as $n \sim \infty$, and where $a_0 = 0$. Also let $a_r - a_{r-1} \leqq \eta$, for every value of r.

Consider the limiting sums $\sigma = \sum_{r=1}^{\infty} a_{r-1} m(e_r)$, $\sigma' = \sum_{r=1}^{\infty} a_r m(e_r)$; where, as before, e_r is the measurable set of points at which $a_{r-1} \leq f(x) < a_r$. The difference of the two sums σ, σ' is $\sum_{r=1}^{\infty} (a_r - a_{r-1}) m(e_r)$, which is not greater than $\eta m(E)$. It follows that σ, σ' are both finite, or both infinite. Now let $a_0, a_1, a_2, \dots$ be the end-points of the meshes of a net, fitted on to the infinite linear interval $(0, \infty)$; and consider a system of such nets, for which η has the values in a sequence $\{\eta_n\}$, of diminishing numbers, that converges to zero. If it be assumed that σ, σ' are finite, it is clear that σ has the values in a non-diminishing sequence, and that σ' has the values in a non-increasing sequence, as the successive nets of the system are taken. It follows that σ and σ' have definite limits, as $n \sim \infty$, and these limits must be identical, since $\sigma' - \sigma \sim 0$, as $n \sim \infty$.

It can be shewn that this limit is independent of the particular system of nets employed. For, if σ_1, σ_1' refer to one system $\{D\}$, of nets, and σ_2, σ_2' to a second system $\{D'\}$, we may take nets D, D' of the two systems such that, for D, $\sigma_1' - \sigma_1 \leq \eta m(E)$, and for D', $\sigma_2' - \sigma_2 \leq \eta m(E)$. Now consider the net (D, D') obtained by superimposing the two nets, and let $\bar{\sigma}'$ be the corresponding value of σ', then $\bar{\sigma}'$ is $\leq \sigma_1'$, and $\leq \sigma_2'$; therefore $\bar{\sigma}' - \sigma_1 \leq \eta m(E)$, and $\bar{\sigma}' - \sigma_2$ is $\leq \eta m(E)$, and moreover $\bar{\sigma}' - \sigma_1$, $\bar{\sigma}' - \sigma_2$ are both ≥ 0. It follows that $|\sigma_1 - \sigma_2| \leq \eta m(E)$; and thus the limits of σ_1 and σ_2, as $\eta \sim 0$, are the same.

The limit as $n \sim \infty$, *of* $\sum_{r=1}^{\infty} a_{r-1} m(e_r)$, *which is identical with the limit of* $\sum_{r=1}^{\infty} a_r m(e_r)$, *for the nets of a system* $\{D_n\}$, *defines the value of* $\int_{(E)} f(x)\,dx$, *for an unbounded non-negative function* $f(x)$, *measurable in the measurable set* E, *whenever this limit exists.*

It will be shewn that this definition is completely equivalent to that given in § 386.

Let us suppose that $a_s = N$, a fixed positive number; then if σ be convergent, we have $\sigma = \sum_{r=1}^{r=s} a_{r-1} m(e_r) + R_s$, $\sigma' = \sum_{r=1}^{r=s} a_r m(e_r) + R_s'$, where N can be chosen so large that R_s and R_s' are both less than ϵ.

Assuming that $\int_{(E)} f(x)\,dx$ is the limit of $\sum_{r=1}^{r=s} a_r m(e_r) + R_s'$, for the system of nets, the integer s so varying that $a_s = N$, we see that

$$\int_{(E)} f_N(x)\,dx$$

is the limit of $\sum_{r=1}^{r=s} a_r m(e_r) + R_s''$, where $R_s'' < R_s' < \epsilon$.

Since R_s', R_s'' differ from one another by less than ϵ, it follows that $\int_{(E)} f(x)\,dx$, $\int_{(E)} f_N(x)\,dx$ differ from one another by not more than ϵ; the number N having been chosen sufficiently large. It follows that

$$\lim_{N\sim\infty}\int_{(E)} f_N(x)\,dx = \int_{(E)} f(x)\,dx.$$

Conversely, let $\int_{(E)} f(x)\,dx$ be defined to be $\lim_{N\sim\infty}\int_{(E)} f_N(x)\,dx$; then $\int_{(E)} f(x)\,dx$ is defined as the repeated limit

$$\lim_{N\sim\infty}\lim_{n\sim\infty}\left\{\sum_{r=1}^{r=s} a_{r-1}m(e_r) + N\sum_{r=s+1}^{\infty} m(e_r)\right\},$$

n denoting the order of the net $(a_0, a_1, a_2, \dots)$, and s denoting the integer, dependent on n, for which $a_s = N$. We assume that this limit exists.

The sum $\sum_{r=1}^{r=s} a_{r-1}m(E_r)$ is non-diminishing as n increases, and also as s increases, hence the repeated limit $\lim_{N\sim\infty}\lim_{n\sim\infty}\sum_{r=1}^{r=s} a_{r-1}m(e_r)$ must exist, as a finite number. Denoting $\sum_{r=1}^{r=s} a_{r-1}m(e_r)$ by c_{nN}, we see that c_{nN} is such that $c_{n'N'} \geq c_{nN}$, if $n' \geq n$, $N' \geq N$. If we write $\xi = 1/n$, $\eta = 1/N$, we may regard c_{nN} as a function $F(\xi, \eta)$ of the two variables ξ, η, and the function is monotone (see § 307). In accordance with the theorem proved in § 307, the repeated limits of $F(\xi, \eta)$, as $\xi \sim 0$, $\eta \sim 0$, have one and the same finite, or infinite, value. That $F(\xi, \eta)$ is defined only when ξ, η are the reciprocals of positive integers makes no difference as regards the validity of the theorem. We now see that, if either of the repeated limits $\lim_{n\sim\infty}\lim_{N\sim\infty} c_{nN}$, $\lim_{N\sim\infty}\lim_{n\sim\infty} c_{nN}$ exists, as a finite number, the other exists, and has the same value as the former.

We thus see that $\lim_{n\sim\infty}\lim_{N\sim\infty}\sum_{r=1}^{r=s} a_{r-1}m(e_r)$ has a finite value. It follows that $\sum_{r=1}^{\infty} a_{r-1}m(e_r)$ is convergent, and that

$$\lim_{n\sim\infty}\sum_{r=1}^{\infty} a_{r-1}m(e_r) = \lim_{N\sim\infty}\lim_{n\sim\infty}\sum_{r=1}^{r=s} a_{r-1}m(e_r).$$

Now $N\sum_{r=s+1}^{\infty} m(e_r) < \sum_{r=s+1}^{\infty} a_{r-1}m(e_r)$; hence, if N be sufficiently large, $N\sum_{r=s+1}^{\infty} m(e_r)$ becomes arbitrarily small, and $\lim_{N\sim\infty} N\sum_{r=s+1}^{\infty} m(e_r) = 0$. We thus have, since this limit is independent of n,

$$\int_{(E)} f(x)\,dx = \lim_{n\sim\infty}\lim_{N\sim\infty}\left[\sum_{r=1}^{r=s} a_{r-1}m(e_r)\right] = \lim_{n\sim\infty}\sum_{r=1}^{\infty} a_{r-1}m(e_r),$$

which proves that the definition of $\int_{(E)} f(x)\,dx$, as $\lim_{n\sim\infty}\sum_{r=1}^{\infty} a_{r-1} m(e_r)$, is equivalent to the definition as $\lim_{N\sim\infty}\int_{(E)} f_N(x)\,dx$.

The following theorem has been proved in the course of the above proof:

If $f(x)$ be a summable function which is ≥ 0, *in the set E, then*

$$\lim_{N\sim\infty} N m\left(E^{(N)}\right) = 0,$$

where $E^{(N)}$ is that part of E in which $f(x) \geq N$.

When $f(x)$ is not restricted to be ≥ 0 in E, we see from the definition of $\int_{(E)} f(x)\,dx$ as $\int_{(E)} f^{+}(x)\,dx - \int_{(E)} f^{-}(x)\,dx$, that *the integral may be defined* as the limit of either of the expressions* $\sum_{-\infty}^{\infty} a_{r-1} m(E_r)$, $\sum_{-\infty}^{\infty} a_r m(E_r)$, *for a system of nets* $(\ldots - a_{-2}, - a_{-1}, a_0, a_1, a_2, \ldots)$ *fitted on to the indefinite interval* $(-\infty, \infty)$.

We have also the following theorem:

If $f(x)$ be summable in the measurable set E, and $E^{(N)}$ denote that part of E for which $|f(x)| \geq N$, then $\lim_{N\sim\infty} N m\left(E^{(N)}\right) = 0$.

OTHER DEFINITIONS OF AN INTEGRAL

389. A general theory of integration has been developed by W. H. Young, independently† of the work of Lebesgue, in two memoirs. In the second of these memoirs, the theory there developed is brought into relation with the work of Lebesgue. The definition of W. H. Young is a generalization of that of the upper, and the lower, R-integral, in which the domain is divided into a number of parts, of a special kind. W. H. Young defines the generalized upper and lower integrals of a function by means of a division of the measurable domain of the independent variable into sets of measurable parts. This definition may be formulated as follows:

Let the measurable set E, which is the domain of integration, be divided into a finite, or enumerably infinite, set of measurable components, and let the measure of each component be multiplied by the upper boundary of the function in that component, and the sums of all such products be formed. The generalized upper integral is defined to be the lower limit of that sum, for all possible modes of division of E into components, as above described.

---

* This is the definition given by Lebesgue; see his memoir "Intégrale, Longueur, Aire," *Annali di Mat.* (III^a), vol. VII (1902), p. 258.

† See the paper "On upper and lower integration," *Proc. Lond. Math. Soc.* (2), vol II (1905), p. 52, and also "On the general theory of integration," *Phil. Trans.* vol. CCIV (1905), p. 221.

The generalized lower integral is defined in a similar manner, by employing the lower boundary of the function in each component. If the generalized upper and lower integrals both exist, and have the same finite value, that value is said to be the value of the integral of the function over E.

This definition can be applied, not only to bounded functions, but also to unbounded functions. In the latter case, such* sets only are employed as have finite upper and lower boundaries of the function. It will be shewn in Vol. II that this definition is equivalent to that of Lebesgue.

The definition of an integral has been further generalized by Pierpont, so as to apply to the case in which the set E is non-measurable.

A set H is said to be *measurable relatively to* E (which may be non-measurable) if there exists a measurable set G such that $H \equiv D(G, E)$. Thus H is that part of E which E has in common with a measurable set.

Pierpont's definition† is obtained from that of W. H. Young by considering those systems of sub-division of E into components which are such that each component is measurable relatively to E. In forming the expression, of which the lower boundary is the upper integral, we employ the exterior measure of each component of E, and the same in forming the expression of which the upper boundary is the lower integral. With these changes in the statement of the definition of W. H. Young, we obtain the definition of Pierpont.

Another mode of defining an integral has been developed by W. H. Young. As this depends upon the theory of monotone sequences of functions, it will be referred to in Vol. II, in that connection.

Another definition of an integral has been given by Borel‡.

THE L-INTEGRAL AS THE MEASURE OF A SET OF POINTS

390. It has been shewn, in § 336, that the R-integral of a function expresses the measure of a set of points that is measurable (J). We proceed to shew that the L-integral has a similar relation to a bounded and measurable set of points. This relation may be expressed in the following theorem:

If the summable function $f(x)$, *defined for the bounded and measurable set of points* E, *be such that* $f(x) \geq 0$, *in* E, *then the set of points defined by* $0 \leq y \leq f(x)$, x *in* E, *has for its measure, plane, or* $(p+1)$ *dimensional, according as* E *is a linear, or a* p*-dimensional, set, the value of* $\int_{(E)} f(x)\,dx$.

---

* See Hildebrandt, *Bulletin Amer. Math. Soc.* vol. XXIV (1917), pp. 120–123.

† *Theory of Functions of Real Variables*, vol. II, pp. 343 *et seq.*

‡ *Journal de Math.* (6), vol. VIII (1912), pp. 199–205; also *Leçons sur la théorie des fonctions*, 2nd ed. pp. 217–256; see also Hahn, *Monatshefte fur Math. u. Physik*, vol. XXVI (1915), p. 3. For a criticism of Borel's definition see Lebesgue, *Annales sc. de l'école normale* (3), vol. XXXV (1918), p. 191. See also Borel, *ibid.* (3), vol. XXXVI (1919), p. 71.

The theorem will first be established for the case in which $f(x) = c$, over the set E.

In that case the two, or $(p+1)$ dimensional, measure of the set of points for which x is in E, and $0 \leq y \leq c$, is $cm(E)$. For the sets E, $C(E)$ can be enclosed in sets of intervals, or cells, such that the measure of the set common to the two sets of intervals is $< \eta$. The set, x in E, $0 \leq y \leq c$, and the complementary set, x in $C(E)$, $0 \leq y \leq c$, can consequently be enclosed in sets of rectangles, or $(p+1)$-dimensional cells, such that the measure of the set common to the two sets is $< c\eta$. As η converges to zero, the measure of the set of intervals, or cells, enclosing E, converges to $m(E)$, and the measure of the set of rectangles, or cells, enclosing the set (x in E, $0 \leq y \leq c$), converges to $cm(E)$, which is thus the measure of the set. Therefore, in this case $\int_{(E)} f(x)\,dx$ is equal to the measure of the set, x in E, $0 \leq y \leq f(x)$.

Next, if $E = \sum_{r=1}^{r=m} e_r$, where e_r denotes a measurable set, and $f(x) = c_r$, for $r = 1, 2, 3, \ldots m$, the measure of the set (x in E, $0 \leq y \leq f(x)$) is the sum of the measures of the r sets (x in e_r, $0 \leq y \leq c_r$), which is $\sum_{r=1}^{r=m} c_r m(e_r)$, and is therefore equal to $\int_{(E)} f(x)\,dx$.

If $f(x)$ is any bounded measurable function (≥ 0), the measure of the set (x in E, $0 \leq y \leq f(x)$) is between the measures of the two sets

$$(x \text{ in } E,\ 0 \leq y \leq \phi_\eta(x)),\quad (x \text{ in } E,\ 0 \leq y \leq \psi_\eta(x)),$$

employed in § 385; that is between $\sum_{r=1}^{r=n} a_{r-1} m(e_r)$ and $\sum_{r=1}^{r=n} a_r m(e_r)$, consequently it differs from $\int_{(E)} f(x)\,dx$ by less than $\eta(U-L)$, which converges to 0, as η does so. Therefore the measure of the set is equal to $\int_{(E)} f(x)\,dx$.

If $f(x)$ be a bounded measurable function which has, in E, values of both signs, we see that

$$\int_{(E)} f(x)\,dx = \int_{(E_1)} f(x)\,dx - \int_{(E_2)} \{-f(x)\}\,dx,$$

where E_1, E_2 are the parts of E in which $f(x)$ is positive and negative respectively.

It follows that $\int_{(E)} f(x)\,dx$ *is the excess of the measure of the set*

$$(x \text{ in } E_1,\ 0 \leq y \leq f(x)),$$

over that of the set $\quad (x \text{ in } E_2,\ 0 \leq y \leq -f(x))$.

The value of the L-integral $\int_a^b f(x)\,dx$, where $f(x)$ is a measurable function ≥ 0, representing the measure of the set of points

$$(a \leq x \leq b,\ 0 \leq y \leq f(x)),$$

may be regarded as representing what is, in geometrical language, called the area bounded by the curve $y = f(x)$, the ordinates at a and b, and the x-axis. Similarly $\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$ may be regarded as representing the volume bounded by the surface $y = f(x^{(1)}, x^{(2)})$, the plane $y = 0$ and the planes $x^{(1)} = a^{(1)}$, $x^{(1)} = b^{(1)}$, $x^{(2)} = a^{(2)}$, $x^{(2)} = b^{(2)}$. Accordingly, such a measure of an area, or of a volume, may exist, in accordance with the definition of Lebesgue, in cases in which it does not exist in accordance with the definition of Jordan, because a set of points which is measurable is not necessarily measurable (J).

The L-integral $\int_a^b f(x)\,dx$, when $f(x)$ has both signs, may be regarded as expressing the excess of the area above the x-axis over the area below it. If $f(x) \geq 0$, in (a, b), and is summable, but unbounded in the interval, the L-integral is the limit, as $N \sim \infty$, of the area defined by $a \leq x \leq b$, $0 \leq y \leq f_N(x)$, the function $f_N(x)$ being defined as in § 386.

THE R-INTEGRAL AS AN L-INTEGRAL

391. It will be shewn that an R-integral is also an L-integral, although the converse does not in general hold.

Let $f(x)$ be integrable (R), in an interval, or cell, (a, b). If E denote the set of points at which $f(x) > A$, any point of E which is a point of continuity of $f(x)$ must be an interior point of E: for such a point has a neighbourhood in which, at every point, $f(x) > A$. A point of E, at which $f(x)$ is discontinuous, need not be an interior point of E, but all such points belong to a set of zero measure, since $f(x)$ is integrable (R). Therefore E consists in general of an open set together with a set of measure zero; and, since both these parts of E are measurable, it follows that E is measurable. Since A is arbitrary, it has thus been shewn that $f(x)$ is a measurable function. Since it is bounded in (a, b), it is therefore summable; and thus has an L-integral. The measure (J) of a set is identical with its measure, when both exist; and therefore the R-integral of $f(x)$ has the same value as the L-integral.

If $f(x)$ is measurable, and bounded, but not integrable (R), it is easily seen that $\overline{\int_a^b} f(x)\,dx \geq \int_a^b f(x)\,dx \geq \underline{\int_a^b} f(x)\,dx$. For, in any mesh δ of a net, the L-integral of $f(x)$ lies between $\delta U(\delta)$ and $\delta L(\delta)$.

THE LEBESGUE INTEGRAL AS A FUNCTION OF A SET OF POINTS

392. It has been pointed out in § 208 that, if a class of sets of points E be assigned, and if by any rule, or set of rules, a definite number is given, corresponding to each set E of the given class, the set of all such numbers may be regarded as defining a function of E which has these numbers for its values.

Let the sets of points E be all measurable sets, and let us assume, for the present, that the class E consists of all measurable sets contained in a measurable domain K which may be bounded or unbounded, but of finite measure. Let a function $f(x)$ be defined for all points of that domain, and let it be taken to be summable in K, whether the function be bounded, or not. It is easily seen that $f(x)$ is summable in any one of the sets E. For $|f(x)|$ is summable in K, and the function $\phi_E(x)$ which is equal to $|f(x)|$ in E, and to 0, in $C(E)$, the complement of E relatively to K, is such that its integral over K is less than that of $|f(x)|$, and is therefore finite. The integral of $\phi_E(x)$ over K is the same as over E; therefore $|f(x)|$ is summable in E, and consequently $f(x)$ is also summable in E.

The value of $\int_{(E)} f(x)\,dx$ may be regarded as defining a function $\phi(E)$ of the measurable set E. In the particular case in which $f(x) = 1$, at all points of K, the L-integral is $m(E)$, which, as has been shewn in § 130, is a completely additive function.

The function $\phi(E) \equiv \int_{(E)} f(x)\,dx$ *converges to zero, as* $m(E) \sim 0$, *uniformly for all sets* E *in the fundamental domain* K.

Let N be an arbitrarily chosen positive number, and let K_N be that part of the fundamental domain K in which $|f(x)| > N$. Let

$$a_0, a_1, a_2, \ldots N, a_{s+1}, a_{s+2}, \ldots,$$

where $a_0 = 0$, $a_s = N$, be a sequence of numbers increasing indefinitely, such that $a_{n+1} - a_n < \epsilon$, for all values of n. The numbers ϵ, N can be so chosen that $\sum_{r=s}^{\infty} a_r m(e_r)$, $\sum_{r=s}^{\infty} a_{r+1} m(e_r)$ both converge to values less than η, where e_r is the set of points for which $a_r \leq |f(x)| < a_{r+1}$. The integral $\int_{(K_N)} |f(x)|\,dx$, which lies between the two sums, is therefore also less than η.

Now $\int_{(E)} |f(x)|\,dx = \int_{(K_{NE})} |f(x)|\,dx + \int_{(E-K_{NE})} |f(x)|\,dx$; where K_{NE} denotes the set of points $D(E, K_N)$, common to E and K_N; that this is the case follows from the definition in § 385.

We see that $\int_{(K_{NE})} |f(x)|\,dx \leq \int_{(K_N)} |f(x)|\,dx < \eta$; hence

$$\int_{(E)} |f(x)|\,dx < \eta + Nm(E - K_{NE}) < \eta + Nm(E) < 2\eta:$$

provided $m(E) < \eta/N$; whatever set E be taken that satisfies this condition. Since η is arbitrary, it follows that $\lim_{m(E)\sim 0} \int_{(E)} |f(x)|\,dx = 0$; and that this convergence is uniform with respect to E.

Since $\left|\int_{(E)} f(x)\,dx\right| \leq \int_{(E)} |f(x)|\,dx$, we see that $\int_{(E)} f(x)\,dx$ converges to 0, as $m(E) \sim 0$, uniformly for all sets E.

A function of E which converges to zero uniformly, as $m(E) \sim 0$, in the fundamental domain K is said to be *an absolutely convergent function of E*.

Accordingly the function $\int_{(E)} f(x)\,dx$ is an absolutely convergent function of E.

If an unbounded set G, of finite measure, be the outer limiting set of a sequence $\{G_n\}$ of bounded measurable sets, each of which is contained in the next, then

$$\int_{(G)} f(x)\,dx = \lim_{n\sim\infty} \int_{(G_n)} f(x)\,dx$$

For, since $m(G) = \lim_{n\sim\infty} m(G_n)$, (see § 134), we have

$$\lim_{n\sim\infty} \int_{(G-G_n)} f(x)\,dx = 0,$$

from which the result follows.

393. Since, in accordance with § 390, $\phi(E)$ is, if $f(x) \geq 0$, the measure of a two-dimensional set, or a $(p+1)$-dimensional set, according as E is a linear set, or a p-dimensional set, it follows from the theorem (§ 130) that the measure of a set is completely additive, that, in the case in which $f(x) \geq 0$, in K, the function $\phi(E)$ is completely additive. In the general case in which $f(x)$ is the difference $f^+(x) - f^-(x)$, of two summable functions, each of which is ≥ 0, by applying the result to each of these functions, we obtain the following theorem:

The function $\phi(E)$ defined as the value of $\int_{(E)} f(x)\,dx$, where E is a measurable set contained in the fundamental interval, or cell, in which $f(x)$ is summable, is a completely additive function of E.

In particular we have

$$\int_{(E_1+E_2)} f(x)\,dx = \int_{(E_1)} f(x)\,dx + \int_{(E_2)} f(x)\,dx.$$

A special case of this theorem is that, if $\delta_1, \delta_2, \ldots \delta_n, \ldots$ be a finite, or an enumerable, sequence of non-overlapping intervals, or cells, contained in the fundamental interval, or cell, and Δ denote their sum, or limiting sum, then

$$\int_{(\Delta)} f(x)\,dx = \sum_{n=1}^{\infty} \int_{(\delta_n)} f(x)\,dx.$$

EQUIVALENT L-INTEGRALS

394. If the values of a function $f(x)$, that is summable over a set E, be altered at points of E belonging to a part E_0, for which $m(E_0) = 0$, the value of $\int_{(E)} f(x)\,dx$ is unaltered. For, in the definition of the integral, the measures of all the sets employed will be unaltered. Moreover, the integral of $f(x)$ over the set E will be the same as over the set $E - E_0$.

Two functions defined for a measurable set E, which have almost everywhere in E, equal values, are said to be *equivalent functions*. Thus all summable and equivalent functions have the same L-integral.

A function that is unbounded in E may be equivalent to a bounded function; for all the points at which the unbounded function is numerically greater than a certain fixed number, may form a set of which the measure is zero.

In some cases, especially in the theory of functions defined by series, it is convenient to admit infinite values of a function at particular points of the set E. If these points form a set E_0 such that $m(E_0) = 0$, we can regard $\int_{(E)} f(x)\,dx$ as existing, and equal to $\int_{(E-E_0)} f(x)\,dx$, whenever this latter integral exists.

If $f(x)$ be $\geqq 0$, in the set E, of measure > 0, and such that $\int_{(E)} f(x)\,dx = 0$, then $f(x)$ is zero almost everywhere in E.

Let E_ϵ be the set of points of E at which $f(x) \geqq \epsilon$, then from the definition of the L-integral, we should have $\int_{(E)} f(x)\,dx \geqq \epsilon m(E_\epsilon)$; it follows that the integral of $f(x)$ cannot vanish unless $m(E_\epsilon) = 0$; and this must be the case for every positive value of ϵ. The set of points at which $f(x) > 0$ is the outer limiting set of the sets $E_{\epsilon_1}, E_{\epsilon_2}, \ldots$, corresponding to a sequence of diminishing values of ϵ, that converges to zero. It follows (§ 131) that the set of points at which $f(x) > 0$, has measure zero.

If the summable function $f(x)$, defined for an interval, or cell, (a, b), be such that its integral over any interval, or cell, whatever, contained in (a, b), is zero, the function $f(x)$ must be zero almost everywhere in (a, b).

Let Δ denote a set of non-overlapping intervals, or cells, containing the points of the set E_1, that part of (a, b) for which $f(x) \geqq 0$. We have then

$$\int_{(E_1)} f(x)\,dx + \int_{(\Delta - E_1)} f(x)\,dx = \int_{(\Delta)} f(x)\,dx = 0.$$

The set Δ can be so chosen that $m(\Delta - E_1)$ is arbitrarily small, and therefore so that the integral of $f(x)$ over $\Delta - E_1$ is arbitrarily small. It follows that $\int_{(E_1)} f(x)\,dx = 0$, and therefore $f(x)$ is zero at almost all points of E_1. Similarly, it can be shewn that $f(x)$ is zero at almost all points of the set of points at which $f(x) \leqq 0$. It follows that $f(x)$ is zero at almost all points of (a, b).

PROPERTIES OF THE LEBESGUE INTEGRAL

395. It has been shewn in § 384 that, if $f_1(x), f_2(x)$ be functions that are measurable in the measurable set E, then their sum $f_1(x) + f_2(x)$ is also measurable in E.

It will now be proved that

$$\int_{(E)} \{f_1(x) + f_2(x)\}\,dx = \int_{(E)} f_1(x)\,dx + \int_{(E)} f_2(x)\,dx.$$

First, let it be assumed that $f_1(x), f_2(x)$ are bounded. If ϵ be an arbitrarily chosen positive number, let $e_r^{(1)}, e_r^{(2)}$ denote the two sets in which

$$r\epsilon \leqq f_1(x) < (r+1)\,\epsilon; \quad r\epsilon \leqq f_2(x) < (r+1)\,\epsilon.$$

Consider the two sums $\sum\limits_{r=\alpha}^{r=\beta} r\epsilon m(e_r^{(1)}), \sum\limits_{r=\alpha}^{r=\beta} r\epsilon m(e_r^{(2)})$, where α, β are integers (positive or negative), such that all the values of $f_1(x), f_2(x)$, in E, are $\geqq \alpha\epsilon$, and $< (\beta+1)\,\epsilon$. If e_{rs} denote the set which is common to $e_r^{(1)}, e_s^{(2)}$, we have

$$\sum_{r=\alpha}^{r=\beta} r\epsilon m(e_r^{(1)}) + \sum_{s=\alpha}^{s=\beta} s\epsilon m(e_s^{(2)}) = \sum_{n=2\alpha}^{n=2\beta} 2n\epsilon \left\{ \sum_{r+s=n} m(e_{rs}) \right\} = \sum_{n=2\alpha}^{n=2\beta} n.2\epsilon m(E_n),$$

where E_n denotes that set of points in which $n\epsilon \leqq f_1(x) + f_2(x) < (n+2)\,\epsilon$. This follows from the fact that $e_r^{(1)} = \sum\limits_{s=\alpha}^{s=\beta} e_{rs}$; and $e_s^{(2)} = \sum\limits_{r=\alpha}^{r=\beta} e_{rs}$. Making ϵ converge to zero, we now see that $\sum\limits_{n=2\alpha}^{n=2\beta} n.2\epsilon m(E_n)$ converges to

$$\int_{(E)} \{f_1(x) + f_2(x)\}\,dx.$$

It thus follows that this integral is the sum of $\int_{(E)} f_1(x)\,dx$ and $\int_{(E)} f_2(x)\,dx$. Next, let one, or both, of the functions $f_1(x), f_2(x)$ be unbounded in E. Let E_N be that set of points in which

$$2N > f_1(x) + f_2(x) > -2N; \quad N > f_1(x) > -N; \quad N > f_2(x) > -N;$$

then $m(E_N) \sim m(E)$, as $N \sim \infty$. We have then

$$\int_{(E_N)} \{f_1(x) + f_2(x)\}\,dx = \int_{(E_N)} f_1(x)\,dx + \int_{(E_N)} f_2(x)\,dx;$$

and since by the theorem of § 392, the integrals of

$$f_1(x) + f_2(x),\ f_1(x),\ f_2(x),$$

taken over the set $E - E_N$, all converge to zero, as $N \sim \infty$, and

$$m(E - E_N) \sim 0,$$

we see that

$$\int_{(E)} \{f_1(x) + f_2(x)\}\,dx = \int_{(E)} f_1(x)\,dx + \int_{(E)} f_2(x)\,dx.$$

396. If $\{f(x)\}^2$ be summable in E, a set of finite measure, so also is $f(x)$, but the converse is not necessarily true. For $\{f(x)\}^2 > |f(x)|$, if $|f(x)| > 1$; and therefore $|f(x)|$ is summable over that part of E in each point of which it is > 1, since $\{f(x)\}^2$ is summable over that part. It follows that $|f(x)|$, and consequently $f(x)$, is summable over E.

If $[f_1(x)]^2$, $[f_2(x)]^2$ are both summable over E, so also is $f_1(x) f_2(x)$; for $|f_1(x) f_2(x)| \leqq \frac{1}{2}[\{f_1(x)\}^2 + \{f_2(x)\}^2]$, and therefore $|f_1(x) f_2(x)|$ is summable over E.

It can be the case that $f_1(x) f_2(x)$ is summable over E, but not $\{f_1(x)\}^2$ or $\{f_2(x)\}^2$.

Assuming that $\{f_1(x)\}^2$, $\{f_2(x)\}^2$ are both summable over E, we have

$$\left|\int_{(E)} f_1(x) f_2(x)\,dx\right| \leqq \int_{(E)} |f_1(x) f_2(x)|\,dx$$
$$\leqq \frac{1}{2}\int_{(E)} \{f_1(x)\}^2\,dx + \frac{1}{2}\int_{(E)} \{f_2(x)\}^2\,dx \quad \ldots(1).$$

Also $\{\lambda\,|f_1(x)| + \mu\,|f_2(x)|\}^2$ is summable over E, as it is the sum of three summable functions; λ, μ denoting constants.

Since

$$\lambda^2 \int_{(E)} \{f_1(x)\}^2\,dx + 2\lambda\mu \int_{(E)} |f_1(x) f_2(x)|\,dx + \mu^2 \int_{(E)} \{f_2(x)\}^2\,dx$$

is essentially positive, we see that

$$\left\{\int_{(E)} f_1(x) f_2(x)\,dx\right\}^2 \leqq \left\{\int_{(E)} |f_1(x) f_2(x)|\,dx\right\}^2$$
$$\leqq \int_{(E)} \{f_1(x)\}^2\,dx \int_{(E)} \{f_2(x)\}^2\,dx \quad \ldots\ldots(2).$$

The inequalities (1) and (2) are of considerable use in questions connected with the convergence of series. The inequality (2), being a generalization of a theorem due to Schwarz, is known as Schwarz's inequality.

397. *If $f_1(x)$ be summable, but not equivalent to a bounded function, it is always possible to determine a summable function $f_2(x)$ so that $f_1(x) f_2(x)$ is not summable*.*

If $a_1, a_2, \dots a_n, \dots$ be a set of positive increasing numbers without an upper limit, and e_n denote the set of points at which $a_n \leq |f_1(x)| < a_{n+1}$: then $m(e_n) > 0$, for an infinite number of values of n. The series $\Sigma\, a_n m(e_n)$ is convergent, being $\leq \int_{(E)} |f_1(x)|\, dx$; let $|f_2(x)| = \frac{1}{na_n m(e_n)}$, for all points of e_n, for each value of n. We have then $\int_{(E)} |f_2(x)|\, dx = \Sigma \frac{1}{na_n}$; and if the numbers a_n be properly chosen (for example $a_n = n$), this series is convergent, and $f_2(x)$ is summable. But the series

$$\Sigma \frac{1}{na_n m(e_n)} a_n m(e_n)$$

being divergent,

$$\int_{(E)} |f_1(x) f_2(x)|\, dx$$

has not a finite value, and therefore $f_2(x)$ is a function such as is required.

398. *If $f_n(x)$ denotes a sequence of measurable functions, defined in the interval, or cell, (a, b), and such that $|f_n(x)|$ is, for every value of n, and for every point x, in the interval, or cell, (a, b), less than a fixed positive number A, then if $\{f_n(x)\}$ converges for each value of x, to the value of the function $f(x)$, that function is summable, and $\int_a^b f(x)\, dx = \lim\limits_{n \sim \infty} \int_a^b f_n(x)\, dx$.*

Since, for any fixed value of x, $|f(x) - f_n(x)| < \epsilon$, provided n be sufficiently great, we see that $|f(x)| < |f_n(x)| + \epsilon < A + \epsilon$, and therefore since ϵ is arbitrary, $|f(x)| \leq A$. Therefore the function $f(x)$ is bounded. Let e_n denote that measurable set of points of (a, b) at which $f_n(x) > c$: then the set e, of points at which $f(x) > c$, is such that each point of e belongs to all the sets $\{e_n\}$, from and after some value of n dependent on x, and therefore, employing a theorem given in § 131, the set e is measurable. It follows that $f(x)$ is a measurable function, as c can be arbitrarily chosen: and since $f(x)$ is bounded, it is summable in (a, b).

Let g_n denote the set of points x, at which $|f(x) - f_n(x)| > \epsilon$; the set g_n is measurable, and $\lim\limits_{n \sim \infty} m(g_n) = 0$. For if this is not the case, there are an indefinitely great number of values of n for which $m(g_n)$ is greater than some fixed positive number α. In accordance with the theorem of § 136, there exists a set of points, of measure $\geq \alpha$, such that each point belongs to an infinite number of the sets g_n. This is inconsistent with the condition that $f_n(x)$ converges to $f(x)$, for each value of x. It has thus been shewn that $m(g_n) \sim 0$, as $n \sim \infty$.

---

* See Lebesgue, *Annales de Toulouse* (3), vol. I (1909), p. 38.

Now

$$\int_a^b \{f(x) - f_n(x)\}\,dx = \int_{g_n} \{f(x) - f_n(x)\}\,dx + \int_{C(g_n)} \{f(x) - f_n(x)\}\,dx$$

The first integral on the right-hand side is numerically less than $2Am(g_n)$, and the second integral is numerically less than ϵl, where l is the measure of the fundamental cell. The integral on the left-hand side is arbitrarily small, for all sufficiently large values of n; and thus the theorem is established.

It may be observed that, in case there is an exceptional set of points. of measure zero, at which the sequence $\{f_n(x)\}$ does not converge, the above theorem still holds. For, in the above proof, we can disregard this set; the integrations being all taken over the complements of this set, the measures of the sets over which they are taken are unaffected.

399. The following theorem is concerned with the L-integral of a function which is, at almost every point of its domain, the limit of a monotone sequence of measurable functions.

Let $\{f_n(x)\}$ be a sequence of non-negative bounded functions, summable in the measurable set of points E, in any number of dimensions, and let it be assumed that, for each value of x, in E, the sequence is monotone and non-diminishing. Let it be further assumed that $\lim\limits_{n\sim\infty}\int_{(E)} f_n(x)\,dx$ has a definite value. Then (1), *the points of E at which the sequence $\{f_n(x)\}$ does not converge to a definite number form a set of points of measure zero; and* (2), *if $f(x)$ denote the limit of the sequence $\{f_n(x)\}$, the integral $\int_{(E)} f(x)\,dx$ exists, and has the value $\lim\limits_{n\sim\infty}\int_{(E)} f_n(x)\,dx$; those points of E at which $f(x)$ is undefined being disregarded. Conversely* (3), *if $\int_{(E)} f(x)\,dx$ exists as a definite number, then $\lim\limits_{n\sim\infty}\int_{(E)} f_n(x)\,dx$ exists, and has the same value as the integral.*

The first two parts of this theorem were given* substantially by Vitali, and by† B. Levi; the proof here given was published‡ by Hobson. Let k denote a positive number, and let the functions $f_n^{(k)}(x)$ be defined by $f_n^{(k)}(x) = f_n(x)$, for values of x for which $f_n(x) \leq k$, and by $f_n^{(k)}(x) = k$, for values of x for which $f_n(x) > k$. Let $f^{(k)}(x)$ be a function such that $f^{(k)}(x) = f(x)$, when $f(x) \leq k$, and $f^{(k)}(x) = k$, when $f(x) > k$. For each value of x, $f_n^{(k)}(x)$ defines a monotone double sequence, that is a monotone function of n and k, if n have the values 1, 2, 3, ..., and k have the values in any monotone increasing sequence of numbers without an upper limit. It is clear that $\int_{(E)} f_n^{(k)}(x)\,dx$, regarded as dependent on n and k, defines a monotone double sequence of non-diminishing numbers. From a property of such double sequences (see § 388), we infer that the two repeated limits

---

* *Rend. del Circ. Mat. di Palermo*, vol. XXIII (1907), p. 137.

† *Rend. dell' Istit. Lombardo*, (2), vol. XXXIX (1906), p. 775.

‡ *Proc. Lond. Math. Soc.* (2), vol. VIII (1910), p. 28.

$$\lim_{k\sim\infty}\lim_{n\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx,\quad \lim_{n\sim\infty}\lim_{k\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx$$

are such as to have the same finite value, if either of them is finite.

Since $f_n^{(k)}(x)$ is bounded, for all values of k and x, we have (see § 398),

$$\lim_{k\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx = \int_{(E)} f_n(x)\,dx,$$

and therefore $\lim\limits_{n\sim\infty}\lim\limits_{k\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx = \lim\limits_{n\sim\infty}\int_{(E)} f_n(x)\,dx$, provided that one of these two limits is assumed to exist.

Also, we have $\lim\limits_{n\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx = \int_{(E)} f^{(k)}(x)\,dx$, since $f_n^{(k)}(x)$ is bounded for all values of n and x. Therefore

$$\lim_{k\sim\infty}\lim_{n\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx = \lim_{k\sim\infty}\int_{(E)} f^{(k)}(x)\,dx,$$

provided one of these two limits exists.

Since $f_n^{(k)}(x)$ is summable, so also is $f^{(k)}(x)$, and therefore also $\lim\limits_{k\sim\infty} f^{(k)}(x)$, or $f(x)$, is measurable. In case $\lim\limits_{k\sim\infty}\int_{(E)} f^{(k)}(x)\,dx$ is finite, its value defines

$$\int_{(E)} f(x)\,dx.$$

First, let it be assumed that $\lim\limits_{n\sim\infty}\int_{(E)} f_n(x)\,dx$ exists. If e_k be the set of points for which $f^{(k)}(x) = k$, we have

$$km(e_k) < \int_{(E)} f^{(k)}(x)\,dx < \lim_{n\sim\infty}\int_{(E)} f_n(x)\,dx.$$

The set of points at which $f(x)$ is indefinitely great is contained in all the sets e_k, hence its measure is

$$\leq m(e_k) < \frac{1}{k}\lim_{n\sim\infty}\int_{(E)} f_n(x)\,dx.$$

Therefore, since k is arbitrarily great, the measure of the set of points at which $f(x) = \infty$ is zero.

Also $\displaystyle\int_{(E)} f(x)\,dx = \lim_{k\sim\infty}\int_{(E)} f^{(k)}(x)\,dx = \lim_{n\sim\infty}\int_{(E)} f_n(x)\,dx.$

Next, if $\int_{(E)} f(x)\,dx$, or $\lim_{k\sim\infty}\lim_{n\sim\infty}\int_{(E)} f_n^{(k)}(x)\,dx$, is assumed to be finite, it will be seen, from what has been proved above, that $\lim_{n\sim\infty}\int_{(E)} f_n(x)\,dx$ exists, and has the value of $\int_{(E)} f(x)\,dx$.

THE LIMITS OF A SEQUENCE OF MEASURABLE FUNCTIONS

400. Let $f_1(x), f_2(x), \ldots f_n(x), \ldots$ be a sequence of measurable functions defined in an interval (a, b), or in a cell of any number of dimensions. Let $v_n(x)$ denote that function which, for each value of x, has for its value the greatest of the numbers $f_1(x), f_2(x), \ldots f_n(x)$. The function $v_n(x)$ is measurable; for the set of points at which $A < v_n(x)$ is that set each point of which belongs to one or more of the measurable sets for which

$$A < f_1(x),\ A < f_2(x),\ \ldots\ A < f_n(x).$$

The functions $v_1(x), v_2(x), \ldots v_n(x), \ldots$ form a sequence which, for any value of x, gives a monotone non-decreasing set of numbers; let $w_1(x)$ denote its limit. The value of $w_1(x)$ is finite, or infinite, for each value of x. Let the function $w_n(x)$ be formed in the same manner from the sequence $v_n(x), v_{n+1}(x), \ldots$, obtained by leaving out the first $n-1$ functions in the sequence $v_1(x), v_2(x), \ldots$. The sets $w_1(x), w_2(x), \ldots$ are all measurable; for the set of points x at which $w_1(x) > A$ is the set of points which belong to an infinite number of the measurable sets for which

$$v_1(x) > A,\ v_2(x) > A,\ \ldots\ v_n(x) > A,\ \ldots;$$

hence $w_1(x)$ is a measurable set, and similarly $w_n(x)$ is measurable for each value of n. The functions $w_1(x), w_2(x), \ldots$ define a monotone non-increasing sequence of numbers for each value of x. The limit of this sequence is $\bar{f}(x)$, where $\bar{f}(x)$ denotes $\overline{\lim}_{n\sim\infty} f_n(x)$. As before, this function $\bar{f}(x)$, being the limit of a monotone sequence of measurable functions, is itself measurable. In a similar manner it can be shewn that $\underline{f}(x) \equiv \underline{\lim}_{n\sim\infty} f_n(x)$ is measurable. It has thus been shewn that:

If $f_1(x), f_2(x), \ldots$ be a sequence of measurable functions of one or more variables, defined in an interval, or a cell, (a, b), the functions $\bar{f}(x), \underline{f}(x)$ are measurable. where $\bar{f}(x), \underline{f}(x)$ denote, for each value of x, the upper and lower limits (finite or infinite) of the sequence of numbers $\{f_n(x)\}$.

If, instead of the upper and lower limits of $\{f_n(x)\}$, their upper and lower boundaries are taken, we have the corresponding theorem:

If $\{f_n(x)\}$ is a sequence of measurable functions of one or more variables, the functions $U(x), L(x)$ are measurable; where $U(x), L(x)$ denote, for each value of x, the upper and lower boundaries of the numbers $\{f_n(x)\}$.

The set of points for which $U(x) > A$, is that set which consists of the points that belong to one or more of the sets for which

$$f_1(x) > A, \quad f_2(x) > A, \ldots.$$

It follows that the set is measurable, and consequently that $U(x)$ is a measurable function. Similarly it can be shewn that $L(x)$ is a measurable function.

THE DERIVATIVES OF A FUNCTION

401. *If $\phi(x)$ be continuous in the linear interval (a, b), the four derivatives $D^+\phi(x)$, $D_+\phi(x)$, $D^-\phi(x)$, $D_-\phi(x)$ (finite or infinite) are measurable functions of x. In particular, if $\phi(x)$ have almost everywhere a differential coefficient $\phi'(x)$, then $\phi'(x)$ is a measurable function of x*

We shall consider the cases of $D^+\phi(x)$ and $D_+\phi(x)$; the cases of the other functions being then deducible.

$D^+\phi(x)$ is the upper limit $\overline{\lim\limits_{h\sim 0}}\, I(x, x+h)$, when $h > 0$, and $I(x, x+h)$ denotes the incrementary ratio, $\{\phi(x+h) - \phi(x)\}/h$. It will be shewn that a sequence of positive values of h, converging steadily to zero, can be determined, such that the upper and lower limits of $I(x, x+h)$, when h has successively the values in the sequence, are for every value of x the same as when h is not restricted to have such values. Let

$$h_1', h_2', \ldots h_n', \ldots$$

be a sequence of diminishing positive numbers, converging to zero, and let $\epsilon_1, \epsilon_2, \ldots \epsilon_n, \ldots$ be another such sequence. Since $I(x, x+h)$ is continuous with respect to (x, h), for all values of x in (a, b), and for all values of h that are > 0, it follows from the uniformity of continuity of $I(x, x+h)$, when h is confined to the interval (h'_{n-1}, h'_n), that this interval can be divided into a definite number r_n of parts, such that

$$|\, I(x, x+h) - I(x, x+h')\, | < \epsilon_n,$$

for every value of x, provided h and h' both lie in one and the same part of the interval (h'_{n+1}, h'_n). Let this sub-division of (h'_{n-1}, h'_n) be made for each value of n, and let $h_1, h_2, h_3, \ldots$ denote the end-points of all the parts of all the intervals. The sequence $h_1, h_2, h_3, \ldots$ converges to zero, and it is a sequence such as satisfies the required condition; for we have $|\, I(x, x+h_m) - I(x, x+h)\, | < \epsilon_s$, provided $h_m \geqq h \geqq h_{m-1}$; the integer s being determinate, corresponding to each value of m. It follows that the upper and lower limits of the sequence

$$I(x, x+h_1), I(x, x+h_2), \ldots I(x, x+h_m), \ldots$$

are identical with those of any other sequence

$$I(x, x+h_1'), I(x, x+h_2'), \ldots I(x, x+h_m'), \ldots,$$

when $$h_1 \geqq h_1' \geqq h_2,\ h_2 \geqq h_2' \geqq h_3, \ldots;$$

and generally $h_m \geqq h_m' \geqq h_{m+1}$; and this is the case for every value of x in (a, b). Therefore the sequence $\{h_n\}$ has the required property.

If we identify the functions $f_n(x)$, of § 400, with the functions $I(x, x + h_n)$, where the sequence h_n is formed as indicated above, we see that the two derivatives $D^+\phi(x)$, $D_+\phi(x)$ are measurable functions, since the functions $\{\phi(x + h_n) - \phi(x)\}/h_n$ are measurable for each value of n.

It follows from the theorem that:

The set of points in which a derivative $D\phi(x)$ of a continuous function is finite is measurable.

For the set is the outer limiting set of the sequence of measurable sets for which $n > D\phi(x) > -n$; for $n = 1, 2, 3, \ldots$.

402. *If E be a set of points in the interval (a, b), at which $Df(x)$, one of the derivatives of a function $f(x)$, continuous in (a, b), has a fixed sign (and is not zero), the set E, when it exists, is unenumerable, and contains a perfect set.*

It will be sufficient to consider the set of points E, at which $D^-f(x) < 0$: the other cases may be treated in a similar manner.

If α be a point of E, another point $\beta\ (> \alpha)$ can be determined such that $f(\beta) - f(\alpha)$ has a negative value $-\lambda$.

Let η be a fixed positive number such that $\eta(\beta - \alpha) = \lambda_1 < \lambda$, and let k be a positive number such that $\lambda_1 + k < \lambda$; we shall consider the continuous function $\phi(x) = f(x) - f(\alpha) + \eta(x - \alpha) + k$, in the interval (α, β), of x. Since $\phi(\alpha)$ is positive, and $\phi(\beta)$ is negative, $\phi(x)$ must have the value 0 at one or more points within (α, β); let ξ be the point nearest to β at which this is the case. Since $\phi(\xi) = 0$, and $\phi(\beta)$ is negative, $\phi(x)$ is negative for all values of x such that $\xi < x \leq \beta$. Since $\phi(x) - \phi(\xi)$ is negative, it follows that $D^+\phi(\xi) \leq 0$, whence we have $D^+f(\xi) \leq -\eta < 0$, and therefore ξ belongs to E.

The number λ_1 being kept fixed, k may vary continuously within the interval $(0, \lambda - \lambda_1)$; and, to each such value of k, there corresponds a single value of ξ, which belongs to E. It follows that the set E contains a single point corresponding to each value of k within the interval $(0, \lambda - \lambda_1)$. Therefore E contains an unenumerable set of points, of the cardinal number of the continuum, and thus contains a perfect set. Accordingly the theorem* has been established.

403. *If $f(x)$ be a function defined in the interval (a, b), and E be the set of points at which one of the derivatives $Df(x)$ has a fixed sign (and is not*

---

* See de la Vallée Poussin's *Cours d'Analyse*, 2nd ed., vol. I, p. 80. The proof there given corresponds to the case $\eta = 0$; but the conclusion there drawn that $D^+\phi(\xi) < 0$, is incorrect This is remedied above by the introduction of the positive number η.

zero), then, if $Df(x)$ is finite (bounded or unbounded) at every point of E, the exterior measure of E is greater than zero.

It will be sufficient to consider the case in which E is the set of points at which $D^+f(x) < 0$; it being assumed that, at no point of E, is

$$D^+f(x) = -\infty.$$

Let it be assumed, if possible, that $m_e(E) = 0$. Let E_n be that part of E, in the points of which $n - 1 \leqq -D^+f(x) < n$, then $m_e(E_n) = 0$. The set E_n may be enclosed in a set Δ_n, of non-overlapping intervals, of total measure ϵ_n; and this may be done for all the sets E_n ($n = 1, 2, 3, \ldots$). The numbers ϵ_n can be so chosen that $\sum\limits_{n=1} n\epsilon_n$ converges to a value which is less than an arbitrarily chosen number ϵ. Let $\Delta_n(x)$ denote that part of Δ_n which is in the interval (a, x), and let

$$\phi(x) = m\{\Delta_1(x)\} + 2m\{\Delta_2(x)\} + \ldots + nm\{\Delta_n(x)\} + \ldots;$$

the function $\phi(x)$ being consequently $< \epsilon$, for all values of x.

Consider the function $\psi(x) = f(x) - \phi(x)$. Since $\phi(x)$ is monotone and non-diminishing, we have $D^+\psi(x) \geq D^+f(x)$, at all points of $C(E)$. At every point x, of E, if x belongs to E_n, the increment $\phi(x+h) - \phi(x)$, for sufficiently small values of h, is at least nh; hence

$$D^+\psi(x) \geqq n + D^+f(x) > 0.$$

It has thus been shewn that $\psi(x)$ is monotone and non-diminishing in (a, b); thus $\psi(\beta) - \psi(\alpha) \geqq 0$, for any two points α, β, when $\beta > \alpha$. Since $\phi(\beta) - \phi(\alpha) < \epsilon$, it follows that $f(\beta) - f(\alpha) > -\epsilon$. But if α is a point of E, a point β can be so chosen that $f(\beta) - f(\alpha)$ is negative, and has a value $-k$; and thus $k < \epsilon$. Since ϵ can be so chosen as to be less than k, the hypothesis that $m_e(E) = 0$ has been shewn to lead to contradiction. It has therefore been shewn that $m_e(E) > 0$. In case $f(x)$ is continuous, since E is then certainly measurable, its measure must be > 0.

INDEFINITE INTEGRALS

404. If $f(x)$ be summable, whether bounded or not, in the linear interval (a, b), it is also summable in the interval (a, x), where $a \leqq x \leqq b$. The integral $\int_a^x f(x)\,dx$ may be regarded either as a function of the upper limit, or as a function of the set of points which constitute the interval (a, x).

If $F(x)$ be a function defined in (a, x), and such that

$$F(x) - F(a) = \int_a^x f(x)\,dx,$$

it is called the *indefinite integral* of $f(x)$, and is determinate, when $f(x)$ is defined, except for an additive constant.

An indefinite integral is continuous, and of bounded variation, in the interval for which it is defined.

Since
$$F(x+h) - F(x) = \int_x^{x+h} f(x)\,dx,$$
$$F(x) - F(x-h) = \int_{x-h}^{x} f(x)\,dx,$$
and since the integrals on the right-hand side converge, in accordance with the theorem of § 392, to zero, as $h \sim 0$, it follows that $F(x)$ is continuous at every point x of (a, b). Again, if (a, b) be divided into any number n of parts (x_{r-1}, x_r), where $x_0 = a$, $x_n = b$, we have
$$\sum_{r=1}^{r=n} |F(x_r) - F(x_{r-1})| = \sum_{r=1}^{r=n} \left| \int_{x_{r-1}}^{x_r} f(x)\,dx \right|$$
$$\leqq \sum_{r=1}^{r=n} \int_{x_{r-1}}^{x_r} |f(x)|\,dx$$
$$\leqq \int_a^b |f(x)|\,dx.$$

Since this inequality holds for every possible set of sub-divisions of (a, b) into parts, it follows that $F(x)$ is of bounded variation. This may also be proved by utilizing the fact that a summable function $f(x)$ is the difference of two non-negative summable functions $f_1(x)$, $f_2(x)$. Then $F(x)$ is the difference of the two monotone functions, $F_1(x)$, $F_2(x)$, which are the indefinite integrals of $f_1(x)$, $f_2(x)$. Then by the theorem of § 244, it follows that $F(x)$ is of bounded variation.

405. *The indefinite integral $F(x)$ has a finite differential coefficient almost everywhere in (a, b).*

Since $F(x)$ is of bounded variation, it is expressible as the difference of two monotone functions, and in accordance with the theorem of § 298, it consequently has a finite differential coefficient almost everywhere in $(a. b)$.

This theorem is included in the following more general theorem*:

If $F(x)$ be the indefinite integral of a summable function $f(x)$, defined in (a, b), $F(x)$ has, almost everywhere in (a, b), a differential coefficient equal to $f(x)$.

The theorem will first be proved in the case in which $f(x)$ is bounded in (a, b). Let $e(x)$ be the part in (a, x) of a measurable set of points e, in (a, b).

The metric density of e has been defined in § 137 as
$$\lim_{h=0} \frac{e(x+h) - e(x-h)}{2h};$$

---

* See Lebesgue, *Leçons sur l'intégration*, p. 124.

since the continuous function $e(x)$ is monotone, it has, for almost all values of x, a differential coefficient, and wherever this differential coefficient exists, its value is identical with the above limit, and therefore with the metric density of $e(x)$. It follows that $\frac{d}{dx}e(x)$ exists, and has the value 1, at almost all points of e, and that it exists, and has the value 0, at almost all points of $C(e)$ (see § 140).

If L, U denote the lower and upper boundaries of $f(x)$ in (a, b), let (L, U) be divided into parts (a_0, a_1), (a_1, a_2), ... (a_{n-1}, a_n), where $a_0 = L$, $a_n = U$, and where $a_r - a_{r-1} < \epsilon$, for all values of r. Let e_r be the set of points at which $a_{r-1} \leqq f(x) < a_r$, except that e_n is the set at which

$$a_{n-1} \leqq f(x) \leqq a_n;$$

and let $e_r(x)$ be the part of e_r in (a, x). Let $\phi_1(x) = a_{r-1}$, at every point of e_r, for all values of r, and let $\phi_2(x) = a_r$, at each point of e_r, for all values of r.

Also let $$F_1(x) = \int_a^x \phi_1(x)\,dx, \quad F_2(x) = \int_a^x \phi_2(x)\,dx;$$

then $$F_1(x) = a_0 m\{e_1(x)\} + a_1 m\{e_2(x)\} + \ldots + a_{n-1} m\{e_n(x)\},$$

$$F_2(x) = a_1 m\{e_1(x)\} + a_2 m\{e_2(x)\} + \ldots + a_n m\{e_n(x)\}.$$

The function $F_1(x)$ has a differential coefficient equal to $\phi_1(x)$ almost everywhere in (a, b). For $m\{e_r(x)\}$ has a differential coefficient almost everywhere equal to 1, or 0, according as x belongs, or does not belong, to e_r. Similarly $F_2(x)$ has a differential coefficient equal to $\phi_2(x)$ almost everywhere.

Also

$$\frac{F_1(x+h) - F_1(x)}{h} \leqq \frac{F(x+h) - F(x)}{h} \leqq \frac{F_2(x+h) - F_2(x)}{h},$$

for positive and for negative values of h, since $F_1(x)$ cannot increase more than $F(x)$, and $F(x)$ cannot increase more than $F_2(x)$, as x increases.

Therefore the four derivatives of $F(x)$ all lie between $\phi_1(x)$, $\phi_2(x)$ inclusively, at each point x which does not belong to that set, of measure zero, at which $F_1(x)$, $F_2(x)$ do not have differential coefficients equal to $\phi_1(x)$, $\phi_2(x)$ respectively. Now $\phi_1(x)$, $\phi_2(x)$ differ from each other, and from $f(x)$, by less than ϵ; therefore, almost everywhere, the four derivatives of $F(x)$ differ from one another by less than ϵ. Taking a sequence of values of ϵ, which converges to zero, we see that $F(x)$ has a differential coefficient equal to $f(x)$, almost everywhere in (a, b).

The theorem has now been established for the case of the indefinite integral of a bounded function.

In order to prove the theorem for the case in which $f(x)$ is unbounded, the following Lemma* will be required:

If $\phi(x)$ *be a continuous monotone non-diminishing function,* $\phi'(x)$ *is summable in* (a, b), *and*

$$\int_a^b \phi'(x)\,dx \leqq \phi(b) - \phi(a).$$

The fact that $\phi'(x)$ does not exist, or is not finite, at points of a set of measure zero, has no effect on the integral, which is taken over that set of points at which $\phi'(x)$ exists, and is finite.

Let $\chi_N(x) = \phi'(x)$, for values of x such that $\phi'(x) \leqq N$, and $\chi_N(x) = N$, when $\phi'(x) > N$. In accordance with the result obtained in § 401, $\phi'(x)$ and $\chi_N(x)$ are measurable functions. The function $\int_a^x \chi_N(x)\,dx$, being an indefinite integral of a bounded function, has a differential coefficient, equal to $\chi_N(x)$, almost everywhere in (a, b). At all points of (a, b) its derivatives are all $\leq N$. Let the points of the set G, at which $\int_a^b \chi_N(x)\,dx$ does not possess a differential coefficient equal to $\chi_N(x)$, be enclosed in a set of intervals Δ, of measure ϵ/N, and let Δ_x be the part of Δ in the segment (a, x).

We shall consider the function $\phi(x) + Nm(\Delta_x) \equiv \psi(x)$. In G, we have $D\psi(x) \geqq N$; and at points not in G, $D\psi(x) \geqq D\phi(x) \geqq \chi_N(x)$, where D denotes any one of the four derivatives. The increase of the function $\psi(x)$, with x, is not less than that of $\int_a^x \chi_N(x)\,dx$; therefore

$$\phi(b) - \phi(a) + \epsilon \geqq \int_a^b \chi_N(x)\,dx.$$

It follows that $\int_a^b \chi_N(x)\,dx$ is bounded, for all values of N, and it is also greater than $Nm(H_N)$, where H_N is that set of points at which $D\phi(x) > N$. We infer that $\lim\limits_{N\sim\infty} m(H_N) = 0$. If E_N is the set complementary to H_N, we have

$$\int_{(E_N)} D\phi(x)\,dx \leqq \int_a^b \chi_N(x)\,dx \leqq \phi(b) - \phi(a) + \epsilon.$$

If $N \sim \infty$, we see that $\int_a^b D\phi(x)\,dx$ is finite, and $\leqq \phi(b) - \phi(a)$.

From this Lemma we deduce that:

If $\phi(x)$ *be a continuous function, of bounded variation in* (a, b), *then* $D\phi(x)$ *is summable, when taken over the set of points of* (a, b) *at which it is*

---

* See de la Vallée Poussin, *Cours d'Analyse*, 2nd ed., vol. I, p. 266.

finite; $D\phi(x)$ denoting any one of the four derivatives of $\phi(x)$. The integral of $D\phi(x)$ has the same value for all the four derivatives. The set of points at which $D\phi(x)$ is not finite has measure zero.

For $\phi(x)$ is the difference of two monotone non-diminishing functions $\phi_1(x)$, $\phi_2(x)$. Since $\phi_1'(x)$, $\phi_2'(x)$ exist almost everywhere in (a, b), and are summable over the set of points at which they exist, we see that $\phi'(x)$ exists almost everywhere, and is summable. The theorem follows from the fact that $D\phi(x) = \phi'(x)$, almost everywhere in (a, b): thus

$$\int_a^b \phi'(x)\,dx = \int_a^b D\phi(x)\,dx.$$

We are now in a position to prove the general theorem, when $f(x)$ is unbounded.

Consider the function $f_N(x)$ which $= f(x)$, when $f(x) \leq N$, and is equal to N, when $f(x) > N$; it being assumed that $f(x)$ is a non-negative function.

Since $\int_a^x f(x)\,dx$ and $\int_a^x f_N(x)\,dx$ are monotone functions, they both have finite differential coefficients almost everywhere, and that of $\int_a^x f_N(x)\,dx$ clearly does not exceed that of $\int_a^x f(x)\,dx$. Since $f_N(x)$ is bounded, the differential coefficient of $\int_a^x f_N(x)\,dx$ is $f_N(x)$, almost everywhere: therefore $f_N(x) \leqq F'(x)$, almost everywhere. Since N is arbitrarily great, it follows that $f(x) \leq F'(x)$, almost everywhere.

Now $\int_a^b f(x)\,dx = F(b) - F(a) \geqq \int_a^b F'(x)\,dx$, as is seen by employing the Lemma proved above, but since $f(x) \leq F'(x)$, we have

$$\int_a^b f(x)\,dx \leqq \int_a^b F'(x)\,dx.$$

It follows that $\int_a^b [F'(x) - f(x)]\,dx = 0$; and since $F'(x) - f(x) \geqq 0$, it then follows that $F'(x) = f(x)$, almost everywhere. The theorem has therefore been established for a non-negative unbounded function.

Since every summable function is the difference of two non-negative summable functions, the theorem is completely established.

It is easily seen that, at every point of continuity of the function $f(x)$, $\int_a^x f(x)\,dx$ has a differential coefficient equal to $f(x)$. For at such a point, if l, u be the lower and upper boundaries of the function in $(x - h, x + h)$,

$$F(x + h) - F(x) \text{ and } F(x) - F(x - h)$$

are between lh and uh inclusively, and since l and u both converge to $f(x)$, as $h \sim 0$, it follows that $F(x)$ has as its differential coefficient $f(x)$. The set of points at which $F(x)$ has $f(x)$ for its differential coefficient consequently includes all points of continuity of $f(x)$, when such points exist.

The important theorem here established, which is due to Lebesgue, throws light on one of the fundamental questions which arise as to the reversibility of the two processes of integration and differentiation; as it asserts that the process of integration of any summable function can be reversed by differentiation, at almost all points of the interval for which the function is defined, and certainly at every point of continuity of the function.

406. *The necessary and sufficient condition that a function, defined for an interval, may be an indefinite integral is that it should be absolutely continuous in the interval.*

To shew that the condition stated in the theorem is necessary, consider

$$F(x) = F(a) + \int_a^x f(x)\, dx.$$

If Δ be any set of non-overlapping intervals such that $m(\Delta) < \eta$, we also have $m(\Delta_1) < \eta$, $m(\Delta_2) < \eta$, where Δ_1 consists of those intervals of Δ for which the variations of $F(x)$ are positive, and Δ_2 those for which they are negative.

The integral $\int_{(\Delta_1)} f(x)\, dx$ is the sum of those variations of $F(x)$ that are positive, and $\int_{(\Delta_2)} f(x)\, dx$ the sum of those that are negative.

Since $\left|\int_{(E)} f(x)\, dx\right| < \epsilon$, provided $m(E) < \eta$, where η is determined when ϵ is given, we see that the sums of the variations in the intervals of Δ_1 and of the absolute variations in the intervals of Δ_2 are $< \epsilon$, for all sets of intervals Δ, such that $m(\Delta) < \eta$. The condition of absolute continuity (see § 218) of $F(x)$ is therefore satisfied. To prove that the condition stated is sufficient, let us assume that it is satisfied by $F(x)$; it has been proved in § 244 that $F(x)$ is of bounded variation in (a, b).

Since $F(x)$ is of bounded variation it has a finite differential coefficient $f(x)$ almost everywhere. It follows from the Lemma in § 405 that, $F(x)$ being the difference of two monotone functions, its differential coefficient $f(x)$ is summable in (a, b). Also $\int_a^x f(x)\, dx$, which has thus been shewn to exist, has a differential coefficient $f(x)$, almost everywhere. Therefore $F(x) - \int_a^x f(x)\, dx \equiv \phi(x)$ has a differential coefficient equal to zero, except in a set G, of measure zero. Enclose G in non-overlapping intervals of a set Δ, where $m(\Delta) < \epsilon$. Each point α, at which $\phi'(x)$ exists, and has the value 0, is such that, if x is in a sufficiently small neighbourhood of α. $|\phi(x) - \phi(\alpha)| < \epsilon\,|x - \alpha|$. A Lebesgue chain (see § 78) may be defined as follows, so as to reach from a to b. To a point α, of G, attach the part of that interval of Δ that is on the right of α; to a point α, not belonging to G, attach an interval (α, α_1) such that $|\phi(x) - \phi(\alpha)| < \epsilon(x - \alpha)$, if $\alpha < x < \alpha_1$. This interval may be defined uniquely, by taking (α, α_1) to be the maximum of all the intervals which satisfy the condition There is one and only one chain from a to x ($\leq b$) composed of the intervals so defined for every point α such that $a \leqq \alpha < x$ (see § 78)

Now $|\phi(x) - \phi(a)|$ cannot exceed the sum of the absolute values of the differences of the functional values at the ends of an interval of the chain from a to x. To this sum those of the intervals that are parts of the intervals of Δ contribute a part that is $< \eta$, dependent on ϵ, since $\phi(x)$ is absolutely continuous. The other intervals of the chain contribute a part $< \epsilon(b - a)$. Hence $|\phi(x) - \phi(a)| < \eta - \epsilon(b - a)$, and since ϵ is arbitrarily small and η, ϵ converge together to zero, we have $\phi(x) = \phi(a)$.

Therefore $F(x) = F(a) + \int_a^x f(x)\,dx$, and thus $F(x)$ is an indefinite integral.

The following theorem has been proved·

If $F(x)$ be an indefinite integral in (a, b), then $F(x) - F(a) = \int_a^x F'(x)\,dx$: the set of points at which $F'(x)$ may not exist being ignored.

It has been shewn in § 218 that the sum and the product of two absolutely continuous functions are absolutely continuous functions.

It follows that:

The sum and the product of two functions which are indefinite integrals in (a, b) are both indefinite integrals in the interval.

407. The necessary and sufficient conditions that $F(x)$ should be an indefinite integral have been stated by Lebesgue in another form:

The necessary and sufficient conditions that $F(x)$ should be an indefinite integral in (a, b) are that (1), *it should be of bounded variation in (a, b), and* (2), *the total variation over any set of points of measure zero should be zero.*

By the total variation of a continuous function $F(x)$ over any set E is meant the limit of the sum of the absolute variations of $F(x)$ in a set of non-overlapping intervals Δ which contains E, as $m(\Delta) \sim m(E)$, whenever such limit exists.

It has been shewn that, if the condition of the former theorem is satisfied, the condition (1) is satisfied. Also the condition of absolute continuity ensures not only that the condition (2) is satisfied, but also that it is satisfied uniformly for all sets of points of measure zero. Thus the conditions (1) and (2) are necessary. That the conditions are sufficient is shewn by the proof in § 406, if we apply (2) to the set G, in which $\phi(x)$ has not a differential coefficient equal to zero.

The above theorem may also be stated in the following[*] form:

If $x = f(t)$, *where* $f(t)$ *is a continuous monotone function of* t, *and* $a = f(t_0)$, *the necessary and sufficient condition that*

$$f(t) - f(t_0) = \int_{t_0}^{t} f'(t)\,dt$$

is that, to every set of points on the t-interval, of measure zero, there shall correspond a set of points on the x-interval, of measure zero.

For the total variation of $f(t)$ over any set of points $G^{(t)}$, on the t-interval, is the lower limit of the measure of a set of non-overlapping intervals on the x-interval which encloses the set $G^{(x)}$ of points that corresponds to $G^{(t)}$; and this lower limit is $m_e(G^{(x)})$. Thus the condition that the total variation of $f(t)$ over a set $G^{(t)}$, of measure zero, should be zero, is equivalent to the condition that $m(G^{(x)})$ should be zero.

A continuous function of bounded variation is an indefinite integral, if both the monotone continuous functions of which it is the difference satisfy the condition of the above theorem.

It will be observed that the functions which are indefinite integrals form a sub-class of the class of functions of bounded variation.

A function which is expressible as the sum of an indefinite integral and of a bounded monotone (not necessarily continuous) function, and which is therefore of bounded variation, has been named by[†] W. H. Young an *upper semi-integral*, or a *lower semi-integral*, according as the monotone function is non-diminishing or non-increasing.

It has been shewn by W. H. Young that an integral is both an upper, and a lower, semi-integral. It is accordingly sufficient, in order that a given function may be an indefinite integral, that it be less than some indefinite integral by a monotone non-diminishing function, and also greater than some indefinite integral by a monotone non-diminishing function.

It can be shewn that the necessary and sufficient conditions that $F(x)$ should be an indefinite R-integral are that it should have bounded derivatives in (a, b), and that a derivative should be continuous almost everywhere.

408. *If* $F_1(x)$, $F_2(x)$ *are indefinite integrals, and* $\overline{F}(x)$ *be the function that is equal to* $F_1(x)$, *for all values of* x *such that* $F_1(x) \geq F_2(x)$, *and to* $F_2(x)$ *when* $F_2(x) > F_1(x)$, *then* $\overline{F}(x)$ *is an indefinite integral.*

---

[*] See Hahn, *Monatshefte für Math. u. Physik*, vol. XXIII (1912), p 163.

[†] *Proc. Lond. Math. Soc.* (2), vol. IX (1911), p 294.

The theorem can be at once extended to the case of any finite number of indefinite integrals $F_1(x), F_2(x), \dots F_n(x)$. The function $\overline{F}(x)$ has at each point x the value of the greatest of the given functions.

To prove the theorem, it will be observed that, in any interval (x_1, x_2), the variation $|\overline{F}(x_2) - \overline{F}(x_1)|$ is equal to that of $F_1(x)$, if $F_1(x_1) \geqq F_2(x_1)$ and $F_1(x_2) \geqq F_2(x_2)$, it is equal to that of $F_2(x)$ if both the inequalities are reversed, and to $|F_1(x_1) - F_2(x_2)|$, or to $|F_1(x_2) - F_2(x_1)|$, if only one of the inequalities is reversed. Considering the case in which

$$F_1(x_1) \geqq F_2(x_1), \quad F_1(x_2) \leqq F_2(x_2):$$

if $F_1(x_1) \geq F_2(x_2)$, we have $0 \leq F_1(x_1) - F_2(x_2) \leq F_1(x_1) - F_1(x_2)$, hence $|F_1(x_1) - F_2(x_2)| \leqq |F_1(x_1) - F_1(x_2)|$; if $F_1(x_1) \leq F_2(x_2)$, we have

$$0 \leqq F_2(x_2) - F_1(x_1) \leqq F_2(x_2) - F_2(x_1),$$

and therefore $\quad |F_1(x_1) - F_2(x_2)| \leqq |F_2(x_1) - F_2(x_2)|$.

A similar remark applies to the case in which

$$F_1(x_1) \leqq F_2(x_1), \quad F_1(x_2) \geq F_2(x_2)$$

Thus the variation of $\overline{F}(x)$ in (x_1, x_2) is in any case not greater than the sum of the separate variations of $F_1(x)$, $F_2(x)$ in the same interval, hence the sum of the absolute variations of $\overline{F}(x)$, in any set of intervals, is not greater than the sum of the absolute variations of $F_1(x)$, and of $F_2(x)$, in the same set of intervals. It then follows that, if $F_1(x)$, $F_2(x)$ are absolutely continuous in (a, b), so also is $\overline{F}(x)$, which is therefore an indefinite integral.

409. The criterion that a function $F(x)$, defined in the interval (a, b), may be the indefinite integral of a function of bounded variation has been obtained* by F. Riesz. It is contained in the following theorem:

The necessary and sufficient condition that the function $F(x)$, defined in the interval (a, b), may be the indefinite integral of a function of bounded variation is that the expression

$$\sum_{r=1}^{r=m-1} \left| \frac{F(x_{r+1}) - F(x_r)}{x_{r+1} - x_r} - \frac{F(x_r) - F(x_{r-1})}{x_r - x_{r-1}} \right|,$$

where $(x_0, x_1, \dots x_m)$ defines a net fitted on to the interval (a, b), and $x_0 = a$, $x_m = b$, should be less than a fixed positive number, independent of the particular net.

That the condition is necessary follows from the fact that

$$\frac{F(x_{r+1}) - F(x_r)}{x_{r+1} - x_r} = \frac{1}{x_{r+1} - x_r} \int_{x_r}^{x_{r+1}} f(x)\, dx = \beta_r;$$

where β_r is some number between the upper and lower boundaries of $f(x)$

---

* *Annales sc. de l'école normale* (3), vol. XXVIII (1911), p. 36.

in the interval (x_r, x_{r+1}). When $f(x)$ is of bounded variation, the expression in the theorem is seen to be bounded for all possible nets. For the proof of the sufficiency of the condition, reference may be made to Riesz's memoir.

THE FUNDAMENTAL THEOREM OF THE INTEGRAL CALCULUS FOR A LEBESGUE INTEGRAL

410. Let $F(x)$ be a continuous function, defined for the interval (a, b), and let it be assumed that $F(x)$ has, at every point of (a, b), a differential coefficient $F'(x)$, that is bounded in (a, b). It has been shewn, in § 401, that $F'(x)$ is a measurable function; being bounded, it is consequently integrable (L).

Let $f_n(x) = \dfrac{F(x + h_n) - F(x)}{h_n}$; where $\{h_n\}$ is a sequence of numbers converging to zero; then $|f_n(x)|$ is bounded, for all values of n and x, as it has the same boundaries as $F'(x)$. By the theorem of § 398, we have

$$\lim_{n\sim\infty} \int_a^x f_n(x)\,dx = \int_a^x F'(x)\,dx.$$

Thus
$$\int_a^x F'(x)\,dx = \lim_{h_n\sim 0} \frac{1}{h_n}\left\{\int_x^{x+h_n} F(x)\,dx - \int_a^{a+h_n} F(x)\,dx\right\}$$
$$= F(x) - F(a).$$

The following theorem has been established.

If $F(x)$ be a function which possesses, at every point of the interval (a, b), a differential coefficient $F'(x)$, bounded in that interval, then $F'(x)$ possesses an L-integral in the interval (a, x), which differs from $F(x)$ by a constant only.

This theorem corresponds to the theorem (B), of § 343. For the R-integral, the theorem is subject to the condition that $F'(x)$ should be integrable (R); but it has here been shewn that the corresponding theorem holds without restriction, when Lebesgue's definition of an integral is employed, so long as the differential coefficient is a bounded function.

411. The following theorem was given* by Lebesgue:

If a function $\phi(x)$, defined in a given interval, be such that one of its derivatives (say $D^+\phi(x)$) is finite at every point, the necessary and sufficient condition that the derivative is summable in the interval is that the given function be of bounded variation in the interval.

The indefinite integral of such a summable derivative is the function of which it is the derivative.

To prove the theorem, we assume that, for each value of x in (a, b), the derivative $D^+\phi(x)$, of the given continuous function $\phi(x)$, has a finite value. Let the unbounded interval $(-\infty, \infty)$ be divided into intervals $(c_\iota, c_{\iota-1})$, where the integer ι has all positive and negative values, including zero, for which we assume $c_0 = 0$; also let $c_{\iota-1} - c_\iota < \epsilon$, for all values of ι. Let e_ι denote that set of points x, in (a, b), for which $c_\iota < D^-\phi(x) \leq c_{\iota-1}$; and arrange the sets e_ι in the order $e_0, e_1, e_{-1}, e_2, e_{-2}, \dots e_n, e_{-n}, \dots$.

---

* *Leçons sur l'intégration*, p. 123. For a discussion of his proof, and for corrections to its original form, see *Atti dei Lincei, Rendiconti*, (5) vol. xv (1) (1906), B. Levi, pp. 433, 551, 674; (5) vol. xv (2) (1906), Lebesgue, p. 3, B. Levi, p. 358. Also (5) vol. xvi (1) (1907), Lebesgue, pp. 92, 283.

Let $k_0, k_1, k_{-1}, k_2, k_{-2}, \dots k_n, k_{-n}, \dots$ be a sequence of positive numbers, all less than 1, so chosen that the limiting sum of

$$k_0 + k_1 \mid c_1 \mid + k_{-1} \mid c_{-1} \mid + \dots + k_n \mid c_n \mid + k_{-n} \mid c_{-n} \mid + \dots$$

is less than ϵ. Let the set e_0 be enclosed within non-overlapping intervals of a set Δ_0, and the complementary set $C(e_0)$ within intervals of a set Δ_0', so that the measure of that set of intervals which is common to Δ_0, Δ_0' does not exceed k_0. Enclose e_1 in a set of intervals Δ_1, and $C(e_0 + e_1)$ in a set of intervals Δ_1': where Δ_1, Δ_1' are both within Δ_0', and have in common a set of intervals of measure not exceeding k_1. Proceeding in this manner, we enclose e_p, where p is positive or negative, in a set of intervals Δ_p, and

$$C(e_0 + e_1 + e_{-1} + \dots + e_p)$$

in a set of intervals Δ_p', so that Δ_p, Δ_p' are both interior to Δ_q', where q immediately precedes p, and such that Δ_p, Δ_p' have in common a set of intervals of measure not exceeding k_p.

We have $m(\Delta_p) - m(e_p) < k_p$; and Δ_p has in common with all the other sets Δ a set of intervals whose measure is $\leq k_0 + k_1 + k_{-1} + \dots + k_p$, say $\leq K_p$.

Since $\sum\limits_p \mid c_p \mid m(\Delta_p) - \sum\limits_p \mid c_p \mid m(e_p)$ does not exceed $\Sigma k_p \mid c_p \mid$, or is $< \epsilon$, we see that $\sum\limits_p \mid c_p \mid m(\Delta_p)$ and $\sum\limits_p \mid c_p \mid m(e_p)$ are either both divergent, or both convergent; and in the latter case the difference of their limiting sums is less than ϵ.

Since $\mid c_p \mid - \epsilon < \mid D^+\phi(x) \mid < \mid c_p \mid + \epsilon$, in the set e_p, we have

$$\sum_p \mid c_p \mid m(e_p) - \epsilon(b - a) < \int_a^b \mid D^+\phi(x) \mid dx < \sum_p \mid c_p \mid m(e_p) + \epsilon(b - a).$$

Hence, if $\sum\limits_p \mid c_p \mid m(e_p)$ is convergent, $\int_a^b \mid D^+\phi(x) \mid dx$ is finite; and conversely. It has now been shewn that the necessary and sufficient condition that

$$\int_a^b \mid D^+\phi(x) \mid dx$$

exists, as a finite number, is that $\sum\limits_p \mid c_p \mid m(\Delta_p)$ should be convergent.

Any point x, in (a, b), belongs to one of the sets e_p; let δ_p be that interval, of the set Δ_p, which contains x. Let $(x, x + h)$ be the longest interval, on the right of x, contained in δ_p, of length not greater than ϵ, and satisfying the condition $c_p - \epsilon \leqq I(x, x+h) \leqq c_{p+1} + \epsilon$. It follows that

$$h\{|c_p| - 2\epsilon\} < |\phi(x+h) - \phi(x)| < h\{|c_p| + 2\epsilon\}.$$

A unique Lebesgue chain, reaching from a to b (see § 78), may be defined by means of these intervals $(x, x+h)$. For all left-hand end-points x_a, of the chain, that belong to one and the same set e_p, let B_p denote the intervals of the corresponding part of the chain. Then

$$\sum_a |\phi(x_a + h) - \phi(x_a)|$$

is between the two numbers $|c_p|\, m(B_p) \pm 2\epsilon m(B_p)$; therefore, for the whole chain,

$$\Sigma\, |\phi(x+h) - \phi(x)|$$

is between the two numbers $\sum_p |c_p|\, m(B_p) \pm 2\epsilon(b-a)$.

Now $\sum_p |c_p|\, m(B_p) \leqq \sum_p |c_p|\, m(\Delta_p)$, and therefore $\sum_p |c_p|\, m(B_p)$ converges if $\int_a^b |D^+\phi(x)|\, dx$ is finite.

It has now been shewn that, if $D^+\phi(x)$ is summable in (a, b),

$$\Sigma\, |\phi(x+h) - \phi(x)|,$$

the total variation of $\phi(x)$ over the Lebesgue chain, is less than

$$\sum_p |c_p|\, m(B_p) + 2\epsilon(b-a),$$

or than

$$\sum_p |c_p|\, m(e_p) + \epsilon + 2\epsilon(b-a),$$

and this is less than

$$\int_a^b |D^+\phi(x)|\, dx + \epsilon + 3\epsilon(b-a).$$

We have also

$$|\phi(b) - \phi(a)| \leqq \Sigma\, |\phi(x+h) - \phi(x)|,$$

and therefore

$$|\phi(b) - \phi(a)| < \int_a^b |D^+\phi(x)|\, dx + \epsilon + 3\epsilon(b-a).$$

Since $|D^+\phi(x)|$ is integrable in any interval (α, β) contained in (a, b), the last inequality may be applied to the interval (α, β). If (a, b) be divided into m parts (α_r, β_r), where $r = 1, 2, 3, \dots m$; we have

$$\sum_{r=1}^{r=m} |\phi(\beta_r) - \phi(\alpha_r)| < \int_a^b |D^+\phi(x)|\, dx + m\epsilon + 3m\epsilon(b-a);$$

and since ϵ is arbitrary, we have

$$\sum_{r=1}^{r=m} |\phi(\beta_r) - \phi(\alpha_r)| \leqq \int_a^b |D^+\phi(x)|\, dx.$$

From this we see that $\phi(x)$ must be of bounded variation in (a, b), in case $\int_a^b D^+\phi(x)\, dx$ exists.

Next, let it be assumed that $\phi(x)$ has bounded variation in (a, b), then $\Sigma \mid \phi(x+h) - \phi(x) \mid$ taken for the intervals of the chain is convergent (see § 243), and therefore $\sum_{p=1}^{\infty} \mid c_p \mid m(B_p)$ converges to a finite number.

Take a fixed value P, of p. All points of Δ_p that are not in an interval of B_p necessarily belong to intervals of Δ_q, where $q \neq p$, and their measure does not exceed K_p; hence

$$\sum_{p=1}^{p=P} \mid c_p \mid m(\Delta_p) - \sum_{p=1}^{p=P} \mid c_p \mid m(B_p) \leq \sum_{p=1}^{p=P} K_p \mid c_p \mid .$$

The numbers k can be subjected to the condition $\sum_{p=1}^{p=P} K_p \mid c_p \mid < \epsilon$, for every value of P; we then have $\sum_{p=1}^{p=P} \mid c_p \mid m(\Delta_p) - \epsilon < \sum_{p=1}^{p=P} \mid c_p \mid m(B_p)$.

It follows that

$$\sum_{p=1}^{p=P} \mid c_p \mid m(e_p) - \epsilon < \sum_{p=1}^{p=P} \mid c_p \mid m(B_p) < \sum_{p=1}^{\infty} \mid c_p \mid m(B_p) - \epsilon,$$

and therefore, if $\sum_{p=1}^{\infty} \mid c_p \mid m(B_p)$ is finite, so also is $\sum_{p=1}^{p=P} \mid c_p \mid m(e_p)$, and this last sum is bounded for all values of P; it therefore follows that $\int_a^b \mid D^+ \phi(x) \mid dx$ exists. It has thus been shewn that it is a sufficient condition for the existence of $\int_a^b D^+ \phi(x)\, dx$, that $\phi(x)$ should be of bounded variation in (a, b).

In any interval of the chain, we have

$$h(c_p - \epsilon) \leq \phi(x+h) - \phi(x) \leq h(c_{p-1} - \epsilon);$$

therefore, for that part of the chain for which $p \leq P$,

$$\sum_{p=1}^{p=P} c_p m(B_p) - \epsilon(b-a) \leqq \sum^{P} \{\phi(x+h) - \phi(x)\} \leq \sum_{p=1}^{p=P} c_p m(B_p) - \epsilon(b-a).$$

It now follows that $\sum^{P} \{\phi(x+h) - \phi(x)\}$ lies between the numbers

$$\sum_{p=1}^{p=P} c_p m(e_p) \pm (b - a + 1)\epsilon.$$

Now P may be so chosen that $\sum_{p=1}^{p=P} c_p m(e_p)$ differs from $\sum_{p=1}^{\infty} c_p m(e_p)$ by less than ϵ; then $\Sigma \{\phi(x+h) - \phi(x)\}$, for the whole chain, is between the two numbers $\sum_{p=1}^{\infty} c_p m(e_p) \pm (b - a + 2)\epsilon$. Therefore $\phi(b) - \phi(a)$ is between

$$\int_a^b D^+ \phi(x)\, dx \pm (b - a + 3)\epsilon;$$

and since ϵ is arbitrary, we have

$$\int_a^b D^+\phi(x)\,dx = \phi(b) - \phi(a);$$

and in general $$\int_a^x D^+\phi(x)\,dx = \phi(x) - \phi(a),$$
if x is in (a, b).

The theorem may be proved for the case in which any one of the other derivatives is employed, in a precisely similar manner. We have then the following theorem:

If $\phi(x)$ be of bounded variation, and have its four derivatives finite at each point, we have

$$\phi(b) - \phi(a) = \int_a^b D^+\phi(x)\,dx = \int_a^b D_+\phi(x)\,dx = \int_a^b D^-\phi(x)\,dx = \int_a^b D_-\phi(x)\,dx.$$

This holds for every interval (α, β) contained in (a, b); in every such interval $\int_\alpha^\beta \{D^+\phi(x) - D_+\phi(x)\}\,dx = 0$; hence (see § 394) we have

$$D^+\phi(x) = D_+\phi(x),$$

almost everywhere. Similarly, by taking the other derivatives in pairs, we see that, almost everywhere, all four derivatives are equal, and thus a finite differential coefficient exists. This is a particular case of the theorem, established in § 298, that any function of bounded variation has, almost everywhere, a finite differential coefficient.

We have, further, the following theorem:

If $\phi(x)$ be of bounded variation, and one of its derivatives $D\phi(x)$ is finite at every point of the interval (a, b), then

$$\int_a^b \phi'(x)\,dx = \phi(b) - \phi(a);$$

the integral being taken over the set of points at which $\phi'(x)$ exists, and is finite. In particular, the theorem holds if $\phi'(x)$ exists at every point, and is finite.

412. The theorem of § 411 may be applied* to obtain a proof of the theorem already established, in § 298, that a function of bounded variation, and in particular, any monotone function, has a finite differential coefficient almost everywhere.

As any function of bounded variation is the difference of two monotone non-diminishing functions, it is sufficient to consider the case of a monotone non-diminishing function $f(x)$. All the derivatives of $f(x)$ are ≥ 0, hence those of $f(x) + kx \equiv \phi(x)$ are all $\geq k$, where k is any chosen positive number. It will be sufficient to prove the theorem for $\phi(x)$. Let $\phi(x) = \xi$, and consider the inverse function $x = F(\xi)$, regarded as defined on the

---

* See W. H. Young, *Quarterly Journal of Math.* vol. XLII (1911), p. 79.

ξ-segment $(\phi(a), \phi(b))$, which is also monotone non-diminishing. Whether $\phi(x)$ is continuous or not, $F(\xi)$ is a continuous function of ξ. All the derivatives of $F(\xi)$ are in the interval $(0, k^{-1})$: and thus, as these are all finite, $F(\xi)$ has a differential coefficient for almost all values of ξ. If K be any fixed number $> k^{-1}$, we have $\Delta x < K\Delta\xi$, and thus we have $\Sigma\Delta x < K\Sigma\Delta\xi$. From this it follows that, to any set of points of measure zero on the ξ-segment, there corresponds a set of points on the x-segment, of measure zero. Therefore $x = F(\xi)$ has a differential coefficient $F'(\xi)$, for almost every value of x; and thus $\phi(x)$ has a differential coefficient almost everywhere. Each point on the ξ-segment at which $F'(\xi) = 0$ can be enclosed in an interval $\Delta\xi$, such that $\Delta x < \epsilon\Delta\xi$, where ϵ is arbitrarily chosen. It follows that, if E be the measure of the set of points on the ξ-segment at which $F'(\xi) = 0$, the measure of the corresponding set on the x-segment is $< \epsilon\{m(E) + \eta\}$, where η is arbitrarily small, and since ϵ is arbitrary, it is zero. Thus the points on the x-axis, for which $F'(\xi) = 0$, and consequently $\phi'(x) = +\infty$, form a set of measure zero. Therefore $\phi'(x)$ exists, and is finite, almost everywhere in (a, b).

413. There remains for consideration the case in which a continuous function $\phi(x)$, of bounded variation, is such that one of its derivatives, say $D^+\phi(x)$, is not everywhere finite.

If $\phi(x)$ is continuous, and $D\phi(x)$ is finite, except at the points of a reducible set G, and if

$$\int_a^x D\phi(x)\,dx$$

exists as an L-integral, then

$$\int_a^x D\phi(x)\,dx = \phi(x) - \phi(a).$$

Consider the closed enumerable set G'. If (a', x) be interior to an interval contiguous to G', then

$$\phi(x) - \phi(a') = \int_{a'}^x D\phi(x)\,dx;$$

and therefore

$$\phi(x) - \int_a^x D\phi(x)\,dx = \phi(a') - \int_a^{a'} D\phi(x)\,dx.$$

It follows that the continuous function

$$\phi(x) - \int_a^x D\phi(x)\,dx$$

is constant in each interval contiguous to G', and therefore, since G' is enumerable, it is constant in the whole interval (a, b), and equal to $\phi(a)$. We have therefore the following theorem:

If $\phi(x)$ be a continuous function, and if one of its derivatives $D\phi(x)$ be *infinite only at the points of a reducible set in* (a, b), *and is summable in* (a, b), *when those points at which it is infinite are ignored, then*

$$\int_a^x D\phi(x)\,dx = \phi(x) - \phi(a), \text{ in } (a, b).$$

The condition that $D\phi(x)$ *is summable may be replaced by the condition that* $\phi(x)$ *should be of bounded variation.*

If the set G were irreducible, but non-dense in (a, b), the set G' would contain a perfect component H. The reasoning given above would shew that

$$\phi(x) - \int_a^x D\phi(x)\,dx$$

is constant in each interval contiguous to H, but it would not follow that it would be constant in (a, b). In fact it may be a continuous function with an everywhere dense set of lines of invariability (see § 269). In this case the relation

$$\int_a^x D\phi(x)\,dx = \phi(x) - \phi(a)$$

does not in general hold good; but we have the following theorem:

If $\phi(x)$ *be a continuous function, with one of its derivatives* $D\phi(x)$ *infinite at the points of a non-dense set* G, *such that* G' *consists of a perfect set* H *and an enumerable set (which may be absent), and if* $D\phi(x)$ *possesses an L-integral in* (a, b), *then*

$$\int_a^x D\phi(x)\,dx = \phi(x) - \phi(a) + U(x) - U(a),$$

where $U(x)$ *is a continuous function with an everywhere dense set of lines of invariability. The condition that* $D\phi(x)$ *is summable may be replaced by the condition that* $\phi(x)$ *should be of bounded variation.*

It has already been shewn, in § 405, that, if $\phi(x)$ be continuous, and of bounded variation, $D\phi(x)$ is summable, when taken over the set of points (necessarily of measure equal to that of the whole interval) at which it is finite.

414. It has been shewn in § 405 that, for a function $f(x)$, which is continuous, and of bounded variation, $f'(x)$ is summable in the interval for which $f(x)$ is defined. The indefinite integral $\int_a^x f'(x)\,dx$ is however not necessarily equal to $f(x) - f(a)$, unless $f(x)$ satisfies the further condition that it should be an indefinite integral. In accordance with § 407, this condition may be expressed in the form that the total variation of $f(x)$ over any set of points of measure zero should be zero.

In the general case of a continuous function of bounded variation which is not subject to this last condition, the relation of the function to the indefinite integral of $f'(x)$ is expressed by the following theorem, due[*] to de la Vallée Poussin:

If $f(x)$ be a continuous function, of bounded variation in the interval (a, b), then

$$f(x) - f(a) = \int_a^x f'(x)\,dx + \overline{V}_1(x) - \overline{V}_2(x),$$

where $\overline{V}_1(x)$ is the total variation of the monotone continuous function $f_1(x)$ over the set of points in (a, x), at which one of its derivatives $Df_1(x)$ is not finite, and $\overline{V}_2(x)$ is similarly defined, relatively to $f_2(x)$; and $f_1(x)$, $f_2(x)$ are two continuous monotone non-diminishing functions whose difference is $f(x)$.

It will be sufficient to assume that $f(x)$ is monotone, either non-increasing or non-diminishing; the result will be obtained by considering the difference of two such functions.

Let $\phi_N(x) = |Df(x)|$, at all points at which $|Df(x)| > N$, and $\phi_N(x) = 0$, when $|Df(x)| \leqq N$; thus $\phi_N(x) - Df(x) \geqq -N$. Let the set of points E, at which $Df(x)$ is not finite, be enclosed in a set Δ, of non-overlapping intervals; the set Δ can be divided into a finite set Δ_1, and an infinite set Δ_2, such that the sum of the variations $\Sigma\{f(x_r) - f(x_{r-1})\}$, taken for the intervals (x_{r-1}, x_r) of Δ_2 is numerically $< \epsilon$; these variations are taken with their proper sign.

Let $V_1(x)$ be the sum of the variations, each with its proper sign, of the function

$$f(x) - \int_a^x \{Df(x) - \phi_N(x)\}\,dx,$$

for the intervals of Δ_1 that lie in (a, x); and let $V_2(x)$ be the sum of the variations of $f(x)$, taken positively, for the intervals of Δ_2 that lie in (a, x); we thus have $V_2(x) \leqq V_2(b) < \epsilon$. If x is within an interval of Δ_1, or of Δ_2, the variation in the part of that interval on the left of x is counted in $V_1(x)$, or $V_2(x)$.

Let us consider the function

$$\chi(x) \equiv f(x) - \int_a^x \{Df(x) - \phi_N(x)\}\,dx - V_1(x) - V_2(x).$$

In an interval of the set Δ_1, $V_1(x)$ and

$$f(x) - \int_a^x \{Df(x) - \phi_N(x)\}\,dx$$

both increase by the same amount, or both diminish by the same amount, and $V_2(x)$ does not vary; thus such an interval is one of invariability of $\chi(x)$. In a finite interval complementary to Δ_1, $V_1(x)$ does not vary, and

$$\int_a^x \phi_N(x)\,dx,\ V_2(x)$$

---

[*] See the *Cours d'Analyse*, 2nd ed., vol. I, p. 269.

do not diminish. The function

$$f(x) - \int_a^x Df(x)\, dx$$

has a derivative which is equal to zero, for almost all values of x; and it follows that the points where the derivative of $\chi(x)$ is negative form a set of measure zero, in case there are any such points.

Moreover, $D\chi(x)$ has nowhere the value $-\infty$, for $\chi(x)$ is the sum of

$$\int_a^x \{\phi_N(x) - Df(x)\}\, dx,$$

of which the derivative is $\geq -N$, and of

$$f(x) - V_1(x) + V_2(x),$$

of which the derivative is finite, except at points in Δ_2, and at such points cannot be negative, as in any interval $(x, x+h)$ contained in an interval of Δ_2 in which $f(x)$ diminishes, $f(x) + V_2(x)$ does not vary, and also $V_1(x)$ does not vary. Since, in an interval complementary to Δ_1, $D\chi(x)$ has nowhere the value $-\infty$, and is certainly ≥ 0 almost everywhere, it follows from the theorem in § 403, that it is nowhere negative, and therefore $\chi(x)$ never diminishes, for its incrementary ratios are all ≥ 0. It follows that $\chi(x)$ is a monotone non-diminishing function; thus $\chi(x) \geq \chi(a)$, or

$$f(x) - f(a) \geq \int_a^x \{Df(x) - \phi_N(x)\}\, dx - V_1(x) + V_2(x).$$

Since this holds for an arbitrary value of ϵ, let $\epsilon \sim 0$, then $V_2(x) \sim 0$, and $V_1(x)$ converges to the total variation of

$$f(x) - \int_a^x \{Df(x) - \phi_N(x)\}\, dx,$$

over the intervals of Δ, each variation having its proper sign. Now let $N \sim \infty$, then

$$\int_a^x \phi_N(x)\, dx \sim 0,$$

and $V_1(x)$ converges to the variation, taken with its proper sign, of

$$f(x) - \int_a^x Df(x)\, dx \text{ in } \Delta;$$

this may be denoted by $\overline{V}_\Delta(x)$. Now let $m(\Delta) \sim 0$, then the variation of $\int_a^x Df(x)\, dx$, in Δ, converges to zero, and $\overline{V}_\Delta(x)$ converges to $\overline{V}(x)$, the total variation, with its proper sign, of $f(x)$ over the set E. We then have

$$f(x) - f(a) \geq \int_a^x Df(x)\, dx + \overline{V}(x);$$

we can now change $f(x)$ into $-f(x)$, when $\overline{V}(x)$ changes to $-\overline{V}(x)$; we then find that

$$f(x) - f(a) \leq \int_a^x Df(x)\, dx + \overline{V}(x).$$

From the two inequalities, we have

$$f(x) - f(a) = \int_a^x Df(x)\,dx + \overline{\Gamma}(x) = \int_a^x f'(x)\,dx - \overline{\Gamma}(x),$$

where $f(x)$ is monotone, either non-diminishing or non-increasing.

When $f(x)$ is of bounded variation, and the difference of two monotone non-diminishing functions, $f(a) + P(x)$, $N(x)$, we have

$$f(x) - f(a) = \int_a^x f'(x)\,dx - \overline{\Gamma}_1(x) - \overline{\Gamma}_2(x).$$

THE TOTAL VARIATION OF AN INDEFINITE INTEGRAL

415. It has been shewn in § 404 that the indefinite integral $F(x)$, of a function $f(x)$, summable in (a, b), has its total variation

$$V_a^b F(x) \leqq \int_a^b |f(x)|\,dx.$$

It can however be proved that

$$V_a^b F(x) = \int_a^b |f(x)|\,dx.$$

Let the set E_1, of points at which $f(x) \geqq 0$, be enclosed in a set of intervals Δ_1, of which the measure exceeds that of E_1 by an arbitrarily small amount; a finite set $\overline{\Delta}_1$ of these intervals can then be chosen such that the difference of $m(\overline{\Delta}_1)$ and $m(E_1)$ is arbitrarily small. Similarly the set E_2, of points at which $f(x) < 0$, can be enclosed in a finite set $\overline{\Delta}_2$, of intervals, so that $m(\overline{\Delta}_2)$ and $m(E_2)$ differ by an arbitrarily small number. Further, Δ_1, Δ_2, can be so chosen that the measure of the common part of $\overline{\Delta}_1$, $\overline{\Delta}_2$ is arbitrarily small.

We have $f(x) = f^+(x) - f^-(x)$, where $f^+(x) = 0$, over E_2, and $f^-(x) = 0$, over E_1. Remove from $\overline{\Delta}_1$, and from $\overline{\Delta}_2$, the intervals of the set $D(\overline{\Delta}_1, \overline{\Delta}_2)$, which they have in common; there remains then a finite set of intervals Δ, consisting of the remaining parts of the intervals of $\overline{\Delta}_1$ and $\overline{\Delta}_2$. The sum of the absolute variations of $F(x)$ over the intervals Δ differs from the sums, added together, of the absolute variations over $\overline{\Delta}_1$, $\overline{\Delta}_2$, by not more than

$$2\int_{D(\overline{\Delta}_1, \overline{\Delta}_2)} |f(x)|\,dx;$$

and this is arbitrarily small. The sum of the absolute variations over the intervals of $\overline{\Delta}_1$ and $\overline{\Delta}_2$ is

$$\Sigma\left|\int_{(\overline{\delta}_1)} f(x)\,dx\right| + \Sigma\left|\int_{(\overline{\delta}_2)} f(x)\,dx\right|,$$

where $\bar{\Delta}_1 = \{\bar{\delta}_1\}$, $\bar{\Delta}_2 = \{\bar{\delta}_2\}$; and this differs by an arbitrarily small number from

$$\int_{(\bar{\Delta}_1)} f^+(x)\,dx + \int_{(\bar{\Delta}_2)} f^-(x)\,dx,$$

which differs by an arbitrarily small amount from $\int_a^b |f(x)|\,dx$. Therefore a set of intervals Δ can be determined such that the sum of the absolute variations of $F(x)$ over them differs by an arbitrarily small amount from

$$\int_a^b |f(x)|\,dx;$$

consequently this integral is, in accordance with the definition of § 243, the total variation of $F(x)$ over (a, b), since that total variation cannot exceed the integral. We have further

$$V_a^x F(x) = \int_a^x |f(x)|\,dx,$$

and thus

$$\frac{d}{dx} V_a^x F(x)$$

exists, and is equal almost everywhere to $|f(x)|$.

The variation of $F(x)$, over any set E, has been defined (see § 252) as the lower boundary of the sum of the variations of $P(x)$ and $N(x)$, $P(x)$ and $-N(x)$ denoting the positive, and the negative, total variation of $F(x)$ in (a, x), taken for a set of intervals (α_n, β_n) which enclose the points of E, when all possible such sets of intervals are taken into account.

We can shew that the variation of $F(x)$ over E is equal to

$$\int_{(E)} |f(x)|\,dx.$$

For the integral of $|f(x)|$ over (α_n, β_n) is the sum of the variations of $P(x)$ and $N(x)$.

The above theorem holds also* in the case of functions of two, or more, variables, defined in a rectangle, or cell. In the two-dimensional case, the total variation of $F(x^{(1)}, x^{(2)})$ over a measurable set E is the lower boundary of

$$\Sigma \left| \{F(\beta_n^{(1)}, \beta_n^{(2)}) - F(\beta_n^{(1)}, \alpha_n^{(2)}) - F(\alpha_n^{(1)}, \beta_n^{(2)}) + F(\alpha_n^{(1)}, \alpha_n^{(2)})\} \right|,$$

where the rectangles enclosing E are denoted by

$$(\alpha_n^{(1)}, \alpha_n^{(2)}; \beta_n^{(1)}, \beta_n^{(2)}).$$

---

* Lebesgue, *Annales sc. de l'école normale* (3), vol. XXVII (1910), p. 383.

THE GENERALIZED INDEFINITE INTEGRAL

416. If $f(x)$ be measurable in a fundamental interval, or cell, in which it is defined, and E denote any measurable set of points in that interval, or cell, $\int_{(E)} f(x)\,dx \equiv \phi(E)$ may be called the generalized indefinite integral of $f(x)$. It has been shewn in § 392 to be a function of E which converges to zero, as $m(E)$ does so, uniformly for all such sets, and it has been shewn to be a completely additive function and also an absolutely continuous function.

Let $E + \delta E$ be a measurable set which contains E, and let $E - \delta E$ be one which is contained in E. A function $\phi(E)$ is said to be continuous, for the particular set E, if $\phi(E + \delta E)$, $\phi(E - \delta E)$ converge to $\phi(E)$, as $m(\delta E)$ converges to zero. This is the case if, for every positive number ϵ,

$$|\phi(E \pm \delta E) - \phi(E)| < \epsilon,$$

provided $m(\delta E) < \eta_\epsilon$; where η_ϵ depends upon ϵ, and converges to zero as ϵ does so.

If, for every value of ϵ, a value of η_ϵ can be so chosen as to be the same for all measurable sets E in the given domain, the function $\phi(E)$ is said to be uniformly continuous in the given domain.

We obtain now the theorem that:

If $f(x)$ be a summable function, defined in a fundamental cell, or interval, $\int_{(E)} f(x)\,dx$ is a uniformly continuous function of E.

For
$$\int_{(E+\delta E)} f(x)\,dx - \int_{(E)} f(x)\,dx = \int_{(\delta E)} f(x)\,dx$$
and
$$\int_{(E)} f(x)\,dx - \int_{(E-\delta E)} f(x)\,dx = \int_{(\delta E)} f(x)\,dx;$$
and the integral on the right-hand side is numerically $< \epsilon$, for all sets δE, of which the measure is less than a number η_ϵ, dependent only on ϵ.

That $\int_a^x f(x)\,dx$, $\int_{(a^{(1)},\,a^{(2)})}^{(x^{(1)},\,x^{(2)})} f(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)})$ are absolutely continuous with respect to x, or $(x^{(1)}, x^{(2)})$, are particular cases of this theorem.

417. The following theorem is the analogue of the property of a continuous function of a single variable, given in § 214, that the function takes every value between its upper and lower limits in the interval for which it is defined:

If $\phi(e)$ denote the generalized indefinite integral $\int_{(e)} f(x)\,dx$, for all measurable sets e, contained in a given bounded measurable set E, for which the summable function $f(x)$ is defined, e can be so determined that $\phi(e)$ has a *prescribed value lying between the upper and lower boundaries of $\phi(e)$ for all sets e, in E.*

In the first place we see that, if E is a measurable set, in any number of dimensions, a component F, of E, can be so determined that $m(F)$ has any prescribed value between 0 and $m(E)$. For if $x^{(1)}$ be one of the coordinates which determine the position of a point in space, the section $E(\xi)$, of E, between a plane $x^{(1)} = \xi$, and a fixed plane $x^{(1)} = a$, where those points of E for which $x^{(1)} < a$ form a set which has a measure less than the prescribed number, is such that its measure $m\{E(\xi)\}$ is a continuous function of ξ, and thus the theorem follows from the theorem of § 214.

We have now $\int_{(e)} f(x)\,dx = \int_{(e_1)} f(x)\,dx + \int_{(e_2)} f(x)\,dx$, where e_1 is a component of E_1, that part of E in which $f(x)$ is positive, and where e_2 is a component of E_2, that part of E in which $f(x)$ is negative. The upper and lower limits of $\int_{(e)} f(x)\,dx$, in E, are clearly $\int_{(E_1)} f(x)\,dx$, and $\int_{(E_2)} f(x)\,dx$, which may be denoted by $m(G_1)$ and $-m(G_2)$, respectively, where G_1, G_2 are sets whose dimensions exceed that of E by unity. Any number C between these two can be expressed by $A - B$, where

$$A < m(G_1),\quad B > m(G_2).$$

The set e_1 can be so determined that $\int_{(e_1)} f(x)\,dx = A$, and e_2 can be so determined that

$$\int_{(e_2)} f(x)\,dx = -B;$$

then if $e = e_1 + e_2$, we have $\int_{(e)} f(x)\,dx = C$. It should be observed that G_1 and G_2 are unbounded in case $f(x)$ is unbounded in E, but they have finite measures.

THE INDEFINITE INTEGRAL OF A FUNCTION OF TWO VARIABLES

418. It should be observed that the indefinite integral

$$F(x^{(1)}, x^{(2)}) \equiv \int_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$$

can be expressed as the difference of two monotone non-diminishing functions, which are also quasi-monotone (see § 255).

For, let $f_1(x^{(1)}, x^{(2)}) = f(x^{(1)}, x^{(2)})$, at the points at which the function f is $\geqq 0$, and let $f_2(x^{(1)}, x^{(2)}) = -f(x^{(1)}, x^{(2)})$, at the points at which f is negative; thus $f(x^{(1)}, x^{(2)}) = f_1(x^{(1)}, x^{(2)}) - f_2(x^{(1)}, x^{(2)})$, and

$$F(x^{(1)}, x^{(2)}) = F_1(x^{(1)}, x^{(2)}) - F_2(x^{(1)}, x^{(2)}),$$

where F_1 and F_2 are the indefinite integrals of f_1 and f_2 respectively.

The two functions $F_1(x^{(1)}, x^{(2)})$, $F_2(x^{(1)}, x^{(2)})$ are both monotone and non-diminishing, in accordance with the definition of § 253. Again, with the notation of § 254,

$$\Delta_{(x^{(1)}, x^{(2)})}^{(\bar{x}^{(1)}, \bar{x}^{(2)})} F_1(x^{(1)}, x^{(2)}) = \int_{(x^{(1)}, x^{(2)})}^{(\bar{x}^{(1)}, \bar{x}^{(2)})} f_1(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

with a similar result for F_2. It follows that the functions F_1, F_2 are quasi-monotone, in accordance with the definition of § 255.

419. If the function $f(x^{(1)}, x^{(2)})$ be summable in the cell

$$(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$$

it has been shewn* by Fubini and Tonelli that the indefinite integral

$$\int_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, x^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

denoted by $F(x^{(1)}, x^{(2)})$, has the property that $\dfrac{\partial^2 F}{\partial x^{(1)} \partial x^{(2)}}$ exists, and has the value $f(x^{(1)}, x^{(2)})$, at almost all points of the cell; and thus that

$$\frac{\partial^2 F}{\partial x^{(1)} \partial x^{(2)}} = \frac{\partial^2 F}{\partial x^{(2)} \partial x^{(1)}} = f(x^{(1)}, x^{(2)}),$$

almost everywhere in the cell. This result has been however deduced† by W. H. Young from the more general theorem which he established, that any function $F(x^{(1)}, x^{(2)})$ that is of bounded variation in the cell, in accordance with the definition in § 254, has the property that $\dfrac{\partial^2 F}{\partial x^{(2)} \partial x^{(1)}}$ exists almost everywhere in the cell. This property we proceed to establish.

It is sufficient to consider the case of a quasi-monotone function $F(x^{(1)}, x^{(2)})$.

Since F is monotone with respect to $x^{(1)}$, $\dfrac{\partial F}{\partial x^{(1)}}$ exists for each fixed value of $x^{(2)}$, for all values of $x^{(1)}$, with the exception of those belonging to a set $S_{(x^{(2)})}$ of linear measure zero. Let $x^{(2)}$ have all rational values in the interval $(a^{(2)}, b^{(2)})$, then all the linear sets $S_{(x^{(2)})}$ taken together form a set S, of linear measure zero. For every value of $x^{(1)}$ that does not belong to S, $\dfrac{\partial F}{\partial x^{(1)}}$ exists for all rational values of $x^{(2)}$.

Consider any value of $x^{(1)}$ that does not belong to S, then

$$\frac{F(x^{(1)} + h, x^{(2)}) - F(x^{(1)}, x^{(2)})}{h}$$

is monotone non-diminishing with respect to $x^{(2)}$, and the two functions $\overline{D}F(x^{(1)}, x^{(2)})$, $\underline{D}F(x^{(1)}, x^{(2)})$ which denote the greatest and least extreme

---

[*] *Rend. del Circ. Mat. di Palermo*, vol. XL (1916), p. 295.

[†] *Comptes Rendus*, Paris, vol. CLXIV (1917), p. 622.

derivatives of $F(x^{(1)}, x^{(2)})$, with respect to $x^{(1)}$, are also monotone non-diminishing with respect to $x^{(2)}$. Consequently, for the fixed value of $x^{(1)}$, they are both continuous functions of $x^{(2)}$, except for an enumerable set of values of $x^{(2)}$. For a value of $x^{(2)}$ for which they are both continuous, each of them is the limit of a sequence obtained by giving $x^{(2)}$ rational values only. But when $x^{(2)}$ is rational, $\overline{D}F(x^{(1)}, x^{(2)})$ and $\underline{D}F(x^{(1)}, x^{(2)})$ are equal, since $\frac{\partial F}{\partial x^{(1)}}$ exists at such points. Hence, for any value of $x^{(2)}$, not belonging to the enumerable set, the two functions $\overline{D}F$, $\underline{D}F$ are equal, as they are the limits of one and the same sequence of numbers; and therefore $\frac{\partial F}{\partial x^{(1)}}$ exists. It thus appears that, for each value of $x^{(1)}$, not belonging to a set S of linear measure zero, $\frac{\partial F}{\partial x^{(1)}}$ exists, for all values of $x^{(2)}$, except those of an enumerable set.

When $x^{(1)}$ has a fixed value not belonging to S, the two functions

$$\overline{D}F(x^{(1)}, x^{(2)}),\quad \underline{D}F(x^{(1)}, x^{(2)}),$$

being monotone functions of $x^{(2)}$, have differential coefficients with respect to $x^{(2)}$, for almost all values of $x^{(2)}$; thus

$$\frac{\overline{D}F(x^{(1)}, x^{(2)}+k) - \overline{D}F(x^{(1)}, x^{(2)})}{k},\quad \frac{\underline{D}F(x^{(1)}, x^{(2)}+k) - \underline{D}F(x^{(1)}, x^{(2)})}{k}$$

both have definite limits, as $k \sim 0$, if $x^{(2)}$ does not belong to a certain set of measure zero. A sequence of values of k, converging to zero, may be so chosen that none of the values of $x^{(2)}+k$ belong to the exceptional set, which consists of the enumerable set for which $\frac{\partial F}{\partial x^{(1)}}$ does not exist, and of the set of measure zero at which the limits are not definite. For this sequence of values of k, $\overline{D}(x^{(1)}, x^{(2)}+k) = \underline{D}(x^{(1)}, x^{(2)}+k)$; and also $\overline{D}F(x^{(1)}, x^{(2)}) = \underline{D}F(x^{(1)}, x^{(2)})$, if $x^{(2)}$ does not belong to the exceptional set; hence the two sequences are identical, and their common limit is also $\lim\limits_{k\sim 0}\frac{1}{k}\left\{\frac{\partial F(x^{(1)}, x^{(2)}+k)}{\partial x^{(1)}} - \frac{\partial F(x^{(1)}, x^{(2)})}{\partial x^{(1)}}\right\}$; and therefore $\frac{\partial^2 F}{\partial x^{(2)}\partial x^{(1)}}$ exists. When all values of $x^{(1)}$ are considered, the exceptional sets all belong to a set of plane measure zero. The following theorem has now been established:

If $F(x^{(1)}, x^{(2)})$ *be a quasi-monotone function, defined in a given cell,*

$$\frac{\partial^2 F(x^{(1)}, x^{(2)})}{\partial x^{(1)}\partial x^{(2)}},\quad \frac{\partial^2 F(x^{(1)}, x^{(2)})}{\partial x^{(2)}\partial x^{(1)}}$$

both exist, almost everywhere in the cell.

If $F(x^{(1)}, x^{(2)})$ be an indefinite integral, it may be expressed (see § 386) as the difference of two functions $F_1(x^{(1)}, x^{(2)})$, $F_2(x^{(1)}, x^{(2)})$ which are the indefinite integrals of two non-negative functions $f_1(x^{(1)}, x^{(2)})$, $f_2(x^{(1)}, x^{(2)})$, whose difference is the function $f(x^{(1)}, x^{(2)})$, of which $F(x^{(1)}, x^{(2)})$ is the indefinite integral.

Consider the quasi-monotone function

$$F_1(x^{(1)}, x^{(2)}) \equiv \int_{a^{(1)}}^{x^{(1)}} \int_{a^{(2)}}^{x^{(2)}} f_1(x^{(1)}, x^{(2)})\, dx^{(1)}.dx^{(2)}. \qquad \text{(see § 255).}$$

For a fixed value of $x^{(2)}$, $\dfrac{\partial F_1}{\partial x^{(1)}}$ exists, and is equal to $\displaystyle\int_{a^{(2)}}^{x^{(2)}} f_1(x^{(1)}, x^{(2)})\, dx^{(2)}$, for almost all values of $x^{(1)}$, and when this exists, its differential coefficient with respect to $x^{(2)}$ exists, and is equal to $f_1(x^{(1)}, x^{(2)})$, for almost all values of $x^{(2)}$. It thus follows from the proof of the foregoing theorem, that $\dfrac{\partial^2 F_1}{\partial x^{(2)} \partial x^{(1)}}$ exists and is equal to $f_1(x^{(1)}, x^{(2)})$ almost everywhere in the cell. As the corresponding result holds for the function F_2, we have the theorem that:

If $F(x^{(1)}, x^{(2)}) \displaystyle\int_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, x^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$, *where* $f(x^{(1)}, x^{(2)})$ *is summable in the cell* $(a^{(1)}, a^{(2)}, b^{(1)}, b^{(2)})$, *then, almost everywhere in the cell,*

$$\frac{\partial^2 F}{\partial x^{(1)} \partial x^{(2)}} = \frac{\partial^2 F}{\partial x^{(2)} \partial x^{(1)}} = f(x^{(1)}, x^{(2)})$$

This is the analogue of the theorem for an integral of a function of a single variable, given in § 405.

419^1. The function $F(x^{(1)}, x^{(2)})$ being defined in the cell

$$(a^{(1)}, a^{(2)};\ b^{(1)}, b^{(2)}),$$

let us consider at any interior point $(x^{(1)}, x^{(2)})$, the ratio

$$\frac{1}{h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}+h,\, x^{(2)}+h)} F(x^{(1)}, x^{(2)})$$

where h is positive and such that $(x^{(1)} + h, x^{(2)} + h)$ is within the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. We may define the two numbers $D^{++} F(x)$, $D_{++} F(x)$, by the relations

$$D^{++} F(x) = \overline{\lim_{h \sim 0}} \frac{1}{h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}+h,\, x^{(2)}+h)} F(x^{(1)}, x^{(2)}),$$

$$D_{++} F(x) = \underline{\lim_{h \sim 0}} \frac{1}{h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}+h,\, x^{(2)}+h)} F(x^{(1)}, x^{(2)}).$$

In a similar manner we can define six numbers

$$D^{+-} F(x),\ D_{+-} F(x),\ D^{-+} F(x),\ D_{-+} F(x),\ D^{--} F(x),\ D_{--} F(x):$$

for example,

$$D^{+-} F(x) = \overline{\lim_{h \sim 0}} \frac{1}{-h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}+h,\, x^{(2)}-h)} F(x),$$

$$D_{-+} F(x) = \underline{\lim_{h \sim 0}} \frac{1}{-h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}-h,\, x^{(2)}+h)} F(x),$$

$$D^{--} F(x) = \overline{\lim} \frac{1}{h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}-h,\, x^{(2)}-h)} F(x).$$

The eight numbers, so defined, may be termed the eight *special derivatives* of $F(x^{(1)}, x^{(2)})$ at the point $(x^{(1)}, x^{(2)})$. When, at a point, the eight special derivatives have one and the same value, finite or infinite, their value is said to define that of the *special differential coefficient* at the point.

In the case of a p-dimensional function, 2^{p+1} special derivatives can be defined in an analogous manner.

If $F(x^{(1)}, x^{(2)})$ be continuous in the cell, the eight special derivatives (finite or infinite) are measurable functions of $(x^{(1)}, x^{(2)})$.

For example, consider $D^{++} F(x)$, we may then denote the incrementary ratio

$$\frac{1}{h^2} \Delta_{(x^{(1)}, x^{(2)})}^{(x^{(1)}+h,\, x^{(2)}+h)} F(x^{(1)}, x^{(2)})$$

by $I(x, x+h)$, where x symbolizes $(x^{(1)}, x^{(2)})$. The proof of the corresponding theorem for functions of a single variable, given in § 401, is then applicable, without essential change, to the present case.

It follows, as in § 401, that the set of points in which a special derivative $DF(x^{(1)}, x^{(2)})$ of a continuous function is finite is measurable

419². It will now be proved that:

If the function $F(x^{(1)}, x^{(2)})$ is continuous and of bounded variation in the cell $(a^{(1)}, a^{(2)};\ b^{(1)}, b^{(2)})$, a special derivative $DF(x^{(1)}, x^{(2)})$ is finite except at the points of a set of plane measure zero

Consider the special derivative $D^{++} F(x^{(1)}, x^{(2)})$, let K denote the total variation of $F(x^{(1)}, x^{(2)})$ in the cell (a, b). Let it be assumed that, if possible, the measure of the set of points L at which $D^{++} F(x^{(1)}, x^{(2)})$ is infinite has a value $\lambda\ (> 0)$. Let the set of cells

$$(a^{(1)}, a^{(2)};\ a^{(1)} + h,\ a^{(2)} + h)$$

all interior to (a, b), in which h has the values in a sequence converging to zero, and

$$\left| \Delta_{(a^{(1)}, a^{(2)})}^{(a^{(1)}+h,\, a^{(2)}+h)} F(x^{(1)}, x^{(2)}) \right| > \frac{2K}{\lambda} h^2,$$

be denoted by C. It is clear that L belongs to the set G, the inner limiting set defined in § 136¹, in connection with the set C. In accordance with Vitali's theorem there exists an enumerable set of cells, all belonging to C, and no two of which have a point in common, such that the sum of their measures is $\geqq \lambda$. Of these cells there are a finite number such that the sum of their measures is $> \frac{1}{2}\lambda$. Let these be the cells

$$(\alpha_1^{(1)}, \alpha_1^{(2)};\ \alpha_1^{(1)} + h_1,\ \alpha_1^{(2)} + h_1), \quad (\alpha_2^{(1)}, \alpha_2^{(2)};\ \alpha_2^{(1)} + h_2,\ \alpha_2^{(2)} + h_2), \ldots$$

$$(\alpha_n^{(1)}, \alpha_n^{(2)};\ \alpha_n^{(1)} + h_n,\ \alpha_n^{(2)} + h_n).$$

We have

$$\sum_{r=1}^{r=n} \left| \Delta_{(\alpha_r^{(1)}, \alpha_r^{(2)})}^{(\alpha_r^{(1)}+h_r,\, \alpha_r^{(2)}+h_r)} F(x^{(1)}, x^{(2)}) \right| > \tfrac{1}{2}\lambda . \frac{2K}{\lambda} > K;$$

and this is not possible, because K is the total variation of $F(x^{(1)}, x^{(2)})$ in (a, b). It follows that the measure of L must be zero.

A special derivative of a continuous function of bounded variation in the cell (a, b) is summable over the cell, the points at which it is infinite being neglected.

If $D(x^{(1)}, x^{(2)})$ be one of the special derivatives, let e_n be the measurable set of points such that $nh \leq |D(x^{(1)}, x^{(2)})| < (n+1)h$, where h is a positive number and n is a positive or negative integer, or zero. Let $n_1, n_2, \ldots$ be the values of the n, for which $m(e_n) > 0$; and consider first the sets $e_{n_1}, e_{n_2}, \ldots e_{n_p}$. Let ϵ be a positive number less than all the numbers $m(e_{n_1}), m(e_{n_2}), \ldots m(e_{n_p})$, let $P = \sum_{\iota=1}^{\iota=p} |\iota n_\iota|$, and let $\eta = \epsilon/P$. By a corollary to Vitali's theorem (§ 136¹) a finite set of squares, C_1, no two of which have a point in common, exists that enclose a part of e_{n_1}, of measure greater than $m(e_{n_1}) - \eta$, such that the measure of C_1 is $< m(e_{n_1})$. The points of e_{n_2} that are exterior to the squares of C_1 form a set of measure $> m(e_{n_2}) - \eta$; and therefore a finite set of squares C_2, no two of which have a point in common, exists, all exterior to C_1, which encloses a part of e_{n_2}, of measure $> m(e_{n_2}) - 2\eta$, and such that

$$m(C_2) < m(e_{n_2}) - \eta.$$

We proceed in this manner to obtain finite sets of squares $C_3, C_4, \ldots C_p$.

Let C_ι' denote the set of squares each of which is interior to one of the squares of C_ι, and such that, if $(\alpha^{(1)}, \alpha^{(2)}), (\beta^{(1)}, \beta^{(2)})$ are its corners,

$$\left| \frac{\Delta_{(\alpha^{(1)}, \alpha^{(2)})}^{(\beta^{(1)}, \beta^{(2)})} F(x^{(1)}, x^{(2)})}{(\beta^{(1)} - \alpha^{(1)})^2} - n_\iota h \right| < h.$$

The measure of the inner limiting set, corresponding to C_ι' is greater than $m(e_{n_\iota}) - \iota\eta$, and less than $m(e_{n_\iota}) - (\iota - 1)\eta$. Therefore an enumerable set C_ι'', of squares exists, no two of which have a point in common, all of which belong to C_ι', and such that their total measure is $> m(e_{n_\iota}) - \iota\eta$, and $< m(e_{n_\iota}) - (\iota - 1)\eta$.

A finite set of squares belonging to C_ι'' exists, whose total measure has a sum s_ι, between $m(e_{n_\iota}) - \iota\eta$ and $m(e_{n_\iota}) - (\iota - 1)\eta$. Let these squares be denoted by $(\alpha_{\iota\gamma}^{(1)}, \alpha_{\iota\gamma}^{(2)}; \beta_{\iota\gamma}^{(1)}, \beta_{\iota\gamma}^{(2)})$, where $\gamma = 1, 2, 3, \ldots \lambda_\iota$; and we consider these sets of squares, for all the values $1, 2, 3, \ldots p$, of ι.

No two of the squares of the complete finite set have a point in common.

If
$$\frac{\Delta_{(\alpha_{\iota\gamma}^{(1)}, \alpha_{\iota\gamma}^{(2)})}^{(\beta_{\iota\gamma}^{(1)}, \beta_{\iota\gamma}^{(2)})} F(x^{(1)}, x^{(2)})}{(\beta_{\iota\gamma}^{(1)} - \alpha_{\iota\gamma}^{(1)})^2} = n_\iota h + \theta_{\iota\gamma} h,$$

where $|\theta_{\iota\gamma}| < 1$, we have

$$\sum_{\gamma=1}^{\gamma=\lambda_\iota} \Delta_{(\alpha_{\iota\gamma}^{(1)}, \alpha_{\iota\gamma}^{(2)})}^{(\beta_{\iota\gamma}^{(1)}, \beta_{\iota\gamma}^{(2)})} F(x^{(1)}, x^{(2)}) = n_\iota h s_\iota + \theta_\iota h s_\iota,$$

where $|\theta_\iota| < 1$. It follows that

$$\sum_{\iota=1}^{\iota=p}\sum_{\gamma=1}^{\gamma=\lambda_\iota} \Delta_{(\alpha_{\iota\gamma}^{(1)},\, \alpha_{\iota\gamma}^{(2)})}^{(\beta_{\iota\gamma}^{(1)},\, \beta_{\iota\gamma}^{(2)})} F(x^{(1)}, x^{(2)}) = h \sum_{\iota=1}^{\iota=p} n_\iota s_\iota + h \sum_{\iota=1}^{\iota=p} \theta_\iota s_\iota .$$

Since $\left|\sum_{\iota=1}^{\iota=p} \theta_\iota s_\iota\right| < m(\Delta)$, the measure of the cell $(a^{(1)}, a^{(2)}, b^{(1)}, b^{(2)})$, and also

$$\left|\sum_{\iota=1}^{\iota=p} n_\iota s_\iota - \sum_{\iota=1}^{\iota=p} n_\iota m(e_{n_\iota})\right| < \eta \sum_{\iota=1}^{\iota=p} |n_\iota| = P\eta,$$

we have

$$\sum_{\iota=1}^{\iota=p}\sum_{\gamma=1}^{\gamma=\lambda_\iota} \Delta_{(\alpha_{\iota\gamma}^{(1)},\, \alpha_{\iota\gamma}^{(2)})}^{(\beta_{\iota\gamma}^{(1)},\, \beta_{\iota\gamma}^{(2)})} F(x^{(1)}, x^{(2)}) = h \sum_{\iota=1}^{\iota=p} n_\iota m(e_{n_\iota}) + \delta h P\eta + \theta h m(\Delta),$$

where $|\delta| < 1$, $|\theta| < 1$.

Since the numerical value of the expression on the left-hand side is not greater than K, and $\eta P = \epsilon$, where ϵ is arbitrary, it follows that

$$\left| h \sum_{\iota=1}^{\iota=p} n_\iota m(e_{n_\iota}) \right| \leq K + hm(\Delta)$$

Since this holds good for every integer p, it follows (see § 388) that $DF(x)$ is summable over Δ.

419³. It will next be shewn that:

If $F(x^{(1)}, x^{(2)})$ is absolutely continuous in Δ, then the integral

$$\int_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} DF(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) = \Delta_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} F(x),$$

where $DF(x^{(1)}, x^{(2)})$ is any one of the special derivatives of $F(x^{(1)}, x^{(2)})$.

This theorem and the preceding one were given* by Vitali for the case of a function of one variable. The proofs are applicable, without essential change, to the case of functions defined in a cell in any number of dimensions.

If ζ be an arbitrarily chosen positive number, since $F(x^{(1)}, x^{(2)})$ is absolutely continuous, the sum of the moduli of the increments

$$\Delta F(x^{(1)}, x^{(2)}),$$

in every set of rectangles, the total measure of which is less than some number $\bar{\eta}$, dependent on ζ, is $< \zeta$. Using the notation employed in the proof of the last theorem, p can be chosen so large and ϵ so small that

$$\left|\sum_{\iota=1}^{\iota=p} s_\iota - \sum_{\iota=1}^{\iota=p} m(e_{n_\iota})\right| < \tfrac{1}{2}\bar{\eta} \text{ and } m(\Delta) - \sum_{\iota=1}^{\iota=p} m(e_{n_\iota}) < \tfrac{1}{2}\bar{\eta}.$$

The points of Δ that are not interior to some square of the set

$$(\alpha_{\iota\gamma}^{(1)}, \alpha_{\iota\gamma}^{(2)};\ \beta_{\iota\gamma}^{(1)}, \beta_{\iota\gamma}^{(2)}),$$

where $\gamma = 1, 2, 3, \ldots \lambda_\iota$, $\iota = 1, 2, 3, \ldots p$, form a finite set of rectangles

$$(\alpha_r^{(1)}, \alpha_r^{(2)};\ \beta_r^{(1)}, \beta_r^{(2)}),$$

---

* *Atti d. Torino*, vol. XLIII (1908), p. 238. The extension to the case of two or more variables is indicated, p. 244.

where $r = 1, 2, 3, \ldots q$, such that

$$\left| \sum_{r=1}^{r=q} \Delta_{(a_r^{(1)},\, a_r^{(2)})}^{(\beta_r^{(1)},\, \beta_r^{(2)})} F\,(x^{(1)}, x^{(2)}) \right| < \zeta.$$

Since

$$\sum_{r=1}^{r=q} \Delta_{(a_r^{(1)},\, a_r^{(2)})}^{(\beta_r^{(1)},\, \beta_r^{(2)})} F\,(x^{(1)}, x^{(2)}) + \sum_{\iota=1}^{\iota=p} \sum_{\gamma=1}^{\gamma=\lambda_\iota} \Delta_{(\alpha_{\iota\gamma}^{(1)},\, a_{\iota\gamma}^{(2)})}^{(\beta_{\iota\gamma}^{(1)},\, \beta_{\iota\gamma}^{(2)})} F\,(x^{(1)}, x^{(2)})$$
$$= \Delta_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} F\,(x^{(1)}, x^{(2)}),$$

we have

$$\left| \Delta_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} F\,(x^{(1)}, x^{(2)}) - h \sum_{\iota=1}^{\iota=p} n_\iota m\,(e_{n_\iota}) \right| < \zeta + \epsilon h + hm\,(\Delta).$$

The number h can be chosen so small, and the number p so large that

$$\left| h \sum_{\iota=1}^{\iota=p} n_\iota m\,(e_{n_\iota}) - \int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} DF\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)}) \right| < \zeta;$$

also h may be taken so small that $h\,(\epsilon + m\,(\Delta)) < \zeta$. We have now

$$\left| \Delta_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} F\,(x^{(1)}, x^{(2)}) - \int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} DF\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)}) \right| < 3\zeta.$$

Since ζ is arbitrarily small, we have

$$\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} DF\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)}) = \Delta_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} F\,(x^{(1)}, x^{(2)}).$$

As $b^{(1)}$, $b^{(2)}$ may be replaced by $x^{(1)}$, $x^{(2)}$, corresponding to any point in Δ, the general theorem has been established.

Since the integral, over any cell contained in Δ, of the difference of any pair of the special derivatives of an absolutely continuous function is zero, it follows (see § 394) that the two special derivatives are equal and finite almost everywhere in Δ. It then follows that:

A function $F\,(x^{(1)}, x^{(2)})$, absolutely continuous in the cell Δ, has a finite special differential coefficient almost everywhere in Δ.

It has been shown in § 416 that the indefinite integral

$$\int_{(a^{(1)},\, a^{(2)})}^{(x^{(1)},\, x^{(2)})} f\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)}),$$

of a function $f\,(x^{(1)}, x^{(2)})$ that is summable in a cell (a, b), is absolutely continuous in the cell. The converse theorem that an absolutely continuous function is the indefinite integral of one of its special derivatives has been proved above. The two results combined give the theorem that:

The necessary and sufficient condition that a function of two variables, defined in a cell Δ, should be the indefinite integral of some function, summable in the cell, is that it should be absolutely continuous in the cell.

The proof given above is applicable, without essential change, to the case of a function of any number of variables. Another very interesting proof of the theorem, dependent upon the theory of functions of sets of points, has been given* by de la Vallée Poussin.

INTEGRATION BY PARTS FOR THE L-INTEGRAL

420. Let the two L-integrals $\int_a^x U(x)\,dx$, $\int_a^x V(x)\,dx$ be denoted by $u(x)$, $v(x)$ respectively. The functions $u(x)$, $v(x)$ are absolutely continuous, and of bounded variation in the linear interval (a, b); moreover, in accordance with the last theorem in § 406, the product uv is an indefinite integral in (a, b). In accordance with the theorem established in § 405, $u(x)$ and $v(x)$ have differential coefficients equal to $U(x)$, $V(x)$ respectively, almost everywhere in (a, b). Let E denote the set of points at which this is the case; then, since $\frac{du}{dx}$, $\frac{dv}{dx}$ are summable in E, $u\frac{dv}{dx}$, $v\frac{du}{dx}$ are also summable in E, since u and v are bounded. We now have

$$\int_{(E)} \frac{d(uv)}{dx}\,dx = \int_{(E)} v\frac{du}{dx}\,dx + \int_{(E)} u\frac{dv}{dx}\,dx;$$

and since $m(E) = b - a$, and uv is an indefinite integral, we may write this equation

$$\Big[uv\Big]_a^b = \int_a^b v\frac{du}{dx}\,dx + \int_a^b u\frac{dv}{dx}\,dx;$$

and this is equivalent to

$$\int_a^b U(x)\left\{\int_a^x V(x)\,dx\right\}dx = \int_a^b U(x)\,dx\int_a^b V(x)\,dx - \int_a^b V(x)\left\{\int_a^x U(x)\,dx\right\}dx$$

which is the general form for integration by parts, for L-integrals.

MEAN VALUE THEOREMS

421. It is frequently of importance to be able to assign upper and lower limits between which the value of a definite integral lies, in cases where the exact determination of the value is not required. Such estimates of the value of a definite integral may frequently be made by the employment of theorems known as mean value theorems; the most important of these will be here investigated.

If $f(x)$ be a bounded function, summable in the measurable set of points E, we have, denoting by U and L the upper, and the lower, boundary of $f(x)$ in E,

$$Lm(E) \leq \int_{(E)} f(x)\,dx \leq Um(E);$$

and therefore $\int_{(E)} f(x)\,dx = Mm(E)$, where M is some number which satisfies the condition $U \geq M \geq L$.

This holds good whatever be the number of dimensions of the set E.

---

* *Intégrales de Lebesgue* (1916), p. 73.

In case $f(x)$ is a continuous function in the linear interval (a, b), there must be some point in (a, b) at which $f(x) = M$. If that point be denoted by $a + \theta(b - a)$, we obtain the following theorem:

If $f(x)$ be continuous in the linear interval (a, b), then

$$\int_a^b f(x)\,dx = (b - a) f\{a + \theta(b - a)\},$$

where θ is some number such that $0 \leqq \theta \leqq 1$.

Next, let $f(x)$ and $\phi(x)$ be summable in the measurable set E, and suppose $f(x)$ is bounded in E, and $\phi(x)$ everywhere $\geqq 0$ in E. We find immediately from the definition of the integral of $f(x)\,\phi(x)$ in E, that

$$L \int_{(E)} \phi(x)\,dx \leqq \int_{(E)} f(x)\,\phi(x)\,dx \leqq U \int_{(E)} \phi(x)\,dx,$$

where U and L are the upper, and the lower, boundary of $f(x)$ in E.

It follows at once that

$$\int_{(E)} f(x)\,\phi(x)\,dx = M_1 \int_{(E)} \phi(x)\,dx,$$

where M_1 is a number such that $U \geqq M_1 \geqq L$.

In case $f(x)$ be a function that is continuous in the linear interval (a, b), we obtain the following theorem:

If $f(x)$ be continuous in the linear interval (a, b), and $\phi(x)$ be a summable function that is $\geqq 0$,

$$\int_a^b f(x)\,\phi(x)\,dx = f\{a + \theta_1(b - a)\} \int_a^b \phi(x)\,dx,$$

where θ_1 is some number such that $0 \leqq \theta_1 \leqq 1$.

This theorem, including also the more general case in which the integration is over any measurable set E in which $f(x)$ is continuous, is known as the First Mean Value Theorem of the Integral Calculus.

An extension to the case in which $\phi(x)$ is not necessarily everywhere of the same sign, but has a finite lower boundary, is obtained by applying the theorem to $\phi(x) + C$, where C is such that $\phi(x) + C \geqq 0$ in (a, b), or in the set E.

For the case of a function of two variables, where $f(x^{(1)}, x^{(2)})$ is continuous in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, and $\phi(x^{(1)}, x^{(2)})$ is summable, and $\geqq 0$, in that cell, the theorem may be written in the form

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\,\phi(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)})$$
$$= f\{a^{(1)} + \theta_1(b^{(1)} - a^{(1)}), a^{(2)} + \theta_2(b^{(2)} - a^{(2)})\} \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \phi(x^{(1)}, x^{(2)})\,d(x^{(1)}, x^{(2)}),$$

where θ_1, θ_2 satisfy the conditions $0 \leqq \theta_1 \leqq 1$, $0 \leqq \theta_2 \leqq 1$.

422. *If the bounded function $f(x)$ be monotone and non-increasing in the linear interval (a, b), and everywhere $\geqq 0$, and if $\phi(x)$ be summable (whether bounded or not) in (a, b), then*

$$\int_a^b f(x)\,\phi(x)\,dx = f(a)\int_a^\xi \phi(x)\,dx,$$

where ξ is some number such that $a \leqq \xi \leqq b$.

Also, if $f(x)$ be monotone and non-diminishing in (a, b), and everywhere $\geqq 0$, then

$$\int_a^b f(x)\,\phi(x)\,dx = f(b)\int_\xi^b \phi(x)\,dx,$$

where ξ is some number such that $a \leqq \xi \leqq b$.

This theorem was first given* by Bonnet, for the case in which $\phi(x)$ has an R-integral, and was applied by him to the theory of Fourier's series.

Another form of the theorem was obtained by Weierstrass, and also by Du Bois-Reymond†, for the case in which $\phi(x)$ has an R-integral, and it is generally known as the Second Mean Value Theorem. When generalized, so that $\phi(x)$ is only restricted to be summable in (a, b), the theorem may be stated as follows:

If $f(x)$ be monotone and bounded in the linear interval (a, b), and if $\phi(x)$ be summable (bounded or unbounded) in that interval, then

$$\int_a^b f(x)\,\phi(x)\,dx = f(a)\int_a^\xi \phi(x)\,dx + f(b)\int_\xi^b \phi(x)\,dx,$$

where ξ is some point in (a, b).

The theorem in this form is deducible immediately from Bonnet's theorem, by writing $f(x) - f(b)$, or $f(x) - f(a)$, in the two cases, instead of $f(x)$. Bonnet's theorem is however not immediately deducible‡ from the second theorem.

It is clear that, since the value of $\int_a^b f(x)\,\phi(x)\,dx$ is unaltered by changing the values of $f(a)$ and $f(b)$, we may, in the above statement, take, instead of $f(a)$, $f(b)$, any two numbers§ A, B which are such that the function $\psi(x)$, defined by $\psi(a) = A$, $\psi(b) = B$, $\psi(x) = f(x)$, for $a < x < b$, is monotone. We have thus the generalized form of the theorem:

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^\xi \phi(x)\,dx + B\int_\xi^b \phi(x)\,dx,$$

where $A \leq f(a+0)$, $B \geq f(b-0)$, if $f(x)$ is non-diminishing, and

$$A \geq f(a+0),\ B \leq f(b-0),$$

if $f(x)$ is non-increasing.

---

* *Mém. Acad. Belg* vol. XXIII (1850), p. 8, also *Liouville's Journal*, vol. XIV (1849), p. 249.

† *Crelle's Journal*, vol. LXIX (1869), p. 81.

‡ This was pointed out by Pringsheim, *Münch. Ber.* vol. XXX (1900), p. 210, where an account of various proofs of the theorem is given.

§ Du Bois-Reymond, *Schlömilch's Zeitschr.* vol. XX (1875), *Hist. Lit. Abtg.* p. 126.

The value of ξ depends in general upon the chosen values of A and B. In this generalized form, the theorem includes Bonnet's theorem as a particular case. For we may take $A = 0$, $B = f(b)$, if $f(x)$ is positive and non-diminishing, or $A = f(a)$, $B = 0$, in case $f(x)$ is positive and non-increasing.

Various proofs* of the theorem of Weierstrass and Du Bois-Reymond† have been published. In these proofs the function $\phi(x)$ has usually been restricted to change its sign only a finite number of times, and sometimes to be differentiable; but a proof free from the former restriction was given by Du Bois-Reymond. A proof has been given by Pringsheim‡ in which $\phi(x)$ is not restricted to be a bounded function, but may be any function such as possesses an absolutely convergent integral, or in certain cases it may have an integral that is not absolutely convergent.

423. The following proof of the second mean value theorem, in its general form, in which the only restriction imposed upon the function $\phi(x)$ is that it is to be summable, whether bounded or unbounded, was given§ by Hobson.

Let $\phi(x)$ be a function which, whether it be bounded or not, is summable in the interval (a, b). Let $f(x)$ be monotone and non-increasing in the interval, and suppose it to have no negative values.

Let ϵ_r be an arbitrarily chosen positive number, less than

$$f(a+0) - f(b-0);$$

and let the function $f_r(x)$ be defined for the interval (a, b) as follows.

An interval (a, x_1) can be so determined that $f(a+0) - f(x) < \epsilon_r$, for $a \leq x < x_1$, and so that $f(a+0) - f(x_1) \geq \epsilon_r$. In case x_1 is a point of continuity of $f(x)$, we shall have $f(a+0) - f(x_1) = \epsilon_r$, but this will not be the case if x_1 is a point of discontinuity. Next, an interval (x_1, x_2) can be so determined that $f(x_1+0) - f(x) < \epsilon_r$, for $x_1 \leq x < x_2$, and that $f(x_1+0) - f(x_2) \geq \epsilon_r$. Proceeding in this manner to determine intervals (x_2, x_3), (x_3, x_4), ..., for some finite value of n, not exceeding

$$\{f(a+0) - f(b-0)\}/\epsilon_r,$$

we must have

$$f(x_{n-1}+0) - f(x) < \epsilon_r, \text{ for } x_{n-1} \leq x < b;$$

we then take x_n to coincide with b.

Let $f_r(x) = f(a+0)$, for $a \leq x < x_1$; let $f_r(x) = f(x_1+0)$, for $x_1 \leq x < x_2$; and in general let $f_r(x) = f(x_s+0)$, for $x_s \leq x < x_{s+1}$. The function $f_r(x)$, so defined, has only a finite number of values in the interval (a, b); it is monotone and non-increasing, and is ≥ 0. Moreover, we have

---

* For example, by Hankel, *Schlömilch's Zeitschr.* vol. XIV, by Meyer, *Math. Annalen*, vol. VI (1873), p. 313; by C. Neumann, *Kreis- Kugel- und Cylinderfunctionen*, Leipzig, 1881, p. 28, by Hölder, *Göttinger Anzeiger*, 1894, p. 519; by Kowalewski, *Math. Annalen*, vol. LX (1905), p. 151.

† *Crelle's Journal*, vol. LXXIX (1875), p. 42. See also Kronecker's *Vorlesungen*, vol. I.

‡ *Münch. Ber.* vol. XXX (1900), p. 218.

§ *Proc. Lond. Math. Soc.* (2), vol. VII (1909), p. 14.

$$0 \leq f_r(x) - f(x) < \epsilon,$$

for every value of x except for the values $a, x_1, x_2, \dots x_{n-1}, b$, at which it is not necessarily the case.

We have now

$$\int_a^b f_r(x)\,\phi(x)\,dx = f(a+0)\int_a^{x_1}\phi(x)\,dx + f(x_1+0)\int_{x_1}^{x_2}\phi(x)\,dx + \dots$$
$$+ f(x_{n-1}+0)\int_{x_{n-1}}^{b}\phi(x)\,dx.$$

Denoting $\int_a^x \phi(x)\,dx$ by $F(x)$, we have

$$\int_a^b f_r(x)\,\phi(x)\,dx = \{f(a+0) - f(x_1+0)\}\,F(x_1)$$
$$+ \{f(x_1+0) - f(x_2+0)\}\,F(x_2) + \dots$$
$$+ \{f(x_{n-2}+0) - f(x_{n-1}+0)\}\,F(x_{n-1}) + f(x_{n-1}+0)\,F(b).$$

Since $f(a+0) - f(x_1+0)$, $f(x_1+0) - f(x_2+0)$, $\dots$, $f(x_{n-1}+0)$ are all ≥ 0, the expression on the right-hand side will be unaltered in value if $F(x_1)$, $F(x_2)$, …… $F(x_{n-1})$, $F(b)$ be all replaced by a definite number N which lies between the greatest and the least of these numbers. The expression then becomes $Nf(a+0)$ Moreover, since $F(x)$ is continuous in the interval (a, b), some value ξ_r, of x, exists such that $N = F(\xi_r)$. It has therefore been proved that

$$\int_a^b f_r(x)\,\phi(x)\,dx = f(a+0)\int_a^{\xi_r}\phi(x)\,dx,$$

where ξ_r is some point in the interval (a, b).

Also $\left|\int_a^b f_r(x)\,\phi(x)\,dx - \int_a^b f(x)\,\phi(x)\,dx\right| < \epsilon_r\int_a^b |\phi(x)|\,dx.$

It follows that

$$\left|\int_a^b f(x)\,\phi(x)\,dx - f(a+0)\int_a^{\xi_r}\phi(x)\,dx\right| < \eta_r,$$

where $\eta_r = \epsilon_r\int_a^b |\phi(x)|\,dx.$

Let $r = 1, 2, 3, \dots$, where $\epsilon_1, \epsilon_2, \epsilon_3, \dots$ form a sequence which converges to zero; then also $\eta_1, \eta_2, \eta_3, \dots$ form a sequence which converges to zero. The set of points $\xi_1, \xi_2, \xi_3, \dots$ has at least one limiting point; and it is clear that the sequence $\{\epsilon_r\}$ may be so chosen by neglecting, if necessary, a part, that the sequence $\{\xi_r\}$ has a single limiting point ξ.

We have then

$$\left|\int_a^b f(x)\,\phi(x)\,dx - f(a+0)\int_a^{\xi}\phi(x)\,dx\right| < \eta_r + f(a+0)\left|\int_{\xi_r}^{\xi}\phi(x)\,dx\right|.$$

If ζ be an arbitrarily chosen positive number, a value r_1, of r, may be so chosen that $\eta_r < \frac{1}{2}\zeta$, and $f(a+0)\left|\int_{\xi_r}^{\bar{\xi}} \phi(x)\,dx\right| < \frac{1}{2}\zeta$, provided $r \geq r_1$. Then we have

$$\left|\int_a^b f(x)\,\phi(x)\,dx - f(a+0)\int_a^{\bar{\xi}} \phi(x)\,dx\right| < \zeta;$$

and therefore, since ζ is arbitrarily small, we must have

$$\int_a^b f(x)\,\phi(x)\,dx = f(a+0)\int_a^{\bar{\xi}} \phi(x)\,dx \quad \ldots\ldots\ldots\ldots\ldots(1).$$

In a precisely similar manner, when $f(x)$ is non-diminishing in (a, b), and is never negative, it may be shewn that

$$\int_a^b f(x)\,\phi(x)\,dx = f(b-0)\int_{\bar{\eta}}^b \phi(x)\,dx \quad \ldots\ldots\ldots\ldots\ldots(2),$$

where $\bar{\eta}$ is some point in (a, b).

In case $f(a) = f(a+0)$, $f(b) = f(b-0)$, these results are equivalent to Bonnet's form of the second mean value theorem.

Next, let $f(x)$ be only restricted to be bounded and monotone in (a, b), but be unrestricted as regards sign. In case $f(x)$ is non-increasing we may apply the theorem (1) to the function $f(x) - f(b-0)$, and we thus have

$$\int_a^b f(x)\,\phi(x)\,dx = f(a+0)\int_a^{\bar{\xi}} \phi(x)\,dx + f(b-0)\int_{\bar{\xi}}^b \phi(x)\,dx.$$

In case $f(x)$ is non-diminishing in (a, b), we may apply the theorem (2) to the function $f(x) - f(a+0)$, and we then have

$$\int_a^b f(x)\,\phi(x)\,dx = f(a+0)\int_a^{\bar{\eta}} \phi(x)\,dx + f(b-0)\int_{\bar{\eta}}^b \phi(x)\,dx.$$

It has thus been shewn that, if $f(x)$ be bounded and monotone in (a, b), and $\phi(x)$ be summable in the same interval,

$$\int_a^b f(x)\,\phi(x)\,dx = f(a+0)\int_a^X \phi(x)\,dx + f(b-0)\int_X^b \phi(x)\,dx.$$

where X is some point in the interval (a, b).

In order to obtain the more general form of this theorem, let A and B be numbers such that $A \geq f(a+0)$, $B \leq f(b-0)$, when $f(x)$ is non-increasing; or else let $A \leq f(a+0)$, $B \geq f(b-0)$, when $f(x)$ is non-diminishing.

Consider an interval $(a-\lambda, b+\lambda)$ which contains (a, b) in its interior, and let $f(x) = A$, for $a - \lambda \leq x < a$; and $f(x) = B$, for $b < x \leq b + \lambda$; the function $f(x)$ being already defined for $a \leq x \leq b$. Let $\phi(x) = 0$, for $a - \lambda \leq x < a$, and for $b < x \leq b + \lambda$, where $\phi(x)$ has already been defined for $a \leq x \leq b$. Now apply the theorem above established, for the interval

$(a - \lambda, b + \lambda)$, for which $f(a - \lambda + 0) = A$, and $f(b + \lambda - 0) = B$. We then have

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^X \phi(x)\,dx + B\int_X^b \phi(x)\,dx,$$

where X is some point in $(a - \lambda, b + \lambda)$, and which clearly lies in (a, b).

The general form of the second mean value theorem may now be stated as follows.

If $f(x)$ *be bounded and monotone in the interval* (a, b), *and if* $\phi(x)$, *whether bounded or not, be summable in* (a, b), *then, if* A, B *be numbers such that*

$$A \geq f(a + 0),\quad B \leq f(b - 0),$$

or

$$A \leq f(a + 0),\quad B \geq f(b - 0),$$

according as $f(x)$ *is non-increasing, or non-diminishing, in* (a, b),

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^X \phi(x)\,dx + B\int_X^b \phi(x)\,dx,$$

where X *is some number in* (a, b). *The number* X *will depend upon the values of* A *and* B. *In particular, we may have* $A = f(a)$, $B = f(b)$, *or also* $A = f(a + 0)$, $B = f(b - 0)$.

In case the function $f(x)$ is never negative in (a, b), we may take $B = 0$, if $f(x)$ is a non-increasing function; and we may take $A = 0$, if $f(x)$ is a non-diminishing function. We obtain thus the following generalized form of Bonnet's theorem

If $f(x)$ *be a bounded monotone function, never negative in* (a, b), *and if* $\phi(x)$ *be any function summable in* (a, b), *then*

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^X \phi(x)\,dx$$

where A *is any number* $\geqq f(a + 0)$, *and* X *is some number, dependent on* A, *in the interval* (a, b), *provided* $f(x)$ *is non-increasing. If* $f(x)$ *is non-diminishing, we have*

$$\int_a^b f(x)\,\phi(x)\,dx = B\int_X^b \phi(x)\,dx$$

where B *is any number* $\geqq f(b - 0)$, *and* X *is some number in* (a, b), *dependent on the value of* B. *In particular, we may take* $A = f(a)$, $B = f(b)$, *in the two cases.*

424. If the function $f(x)$ be of bounded variation in the linear interval (a, b), and M be the upper boundary of $\left|\int_\alpha^\beta \phi(x)\,dx\right|$, for all pairs of values of α, β in the interval (a, b), we may make use of the fact that

$$f(x) = P(x) - N(x),$$

where $P(x)$, $N(x)$ are monotone non-diminishing functions.

We have

$$\int_a^b \{P(x) - P(a)\}\, \phi(x)\, dx = \{P(b) - P(a)\} \int_\xi^b \phi(x)\, dx$$

with a similar equation for $N(x)$. By subtraction, we have

$$\int_a^b \{f(x) - f(a)\}\, \phi(x)\, dx = \{P(b) - P(a)\} \int_\xi^b \phi(x)\, dx$$
$$- \{N(b) - N(a)\} \int_{\xi'}^b \phi(x)\, dx.$$

Hence we have* the following result:

$$\left| \int_a^b f(x)\, \phi(x)\, dx - f(a) \int_a^b \phi(x)\, dx \right| < M \,.\, V_a^b f(x),$$

$V_a^b f(x)$ denoting the total variation in (a, b), where $f(x)$ is of bounded variation in (a, b), and M is the upper boundary of $\left| \int_\alpha^\beta \phi(x)\, dx \right|$, for all intervals (α, β) contained in (a, b).

This result is of considerable use in estimating the value of the integral of the product of a function of bounded variation and a summable function.

425. Let $\phi(x)$ denote a function of any number of variables, that is summable in a bounded measurable set E; and let $f(x)$ denote a function that is bounded and summable in E, and is everywhere ≥ 0. If U and L denote the upper, and the lower, boundary of $f(x)$, in E, let $a_0, a_1, a_2, \ldots a_n$ be a set of decreasing numbers, where $a_0 = A$, a number $\geq U$, and $L \geq a_n \geq 0$. Let e_{r-1} denote that set of points for which $a_{r-1} \geq f(x) > a_r$, for

$$r = 1, 2, 3, \ldots n.$$

Let $f^{(m)}(x) = a_{r-1}$, in the set e_{r-1}, for each value of r; we have then

$$\int_{(E)} f^{(m)}(x)\, \phi(x)\, dx = a_0 \int_{(e_0)} \phi(x)\, dx + a_1 \int_{(e_1)} \phi(x)\, dx + \ldots + a_{n-1} \int_{(e_{n-1})} \phi(x)\, dx$$
$$= (a_0 - a_1) \int_{(e_0)} \phi(x)\, dx + (a_1 - a_2) \int_{(e_0 + e_1)} \phi(x)\, dx$$
$$+ \ldots + (a_{n-2} - a_{n-1}) \int_{(e_0 + e_1 + \ldots + e_{n-2})} \phi(x)\, dx$$
$$+ a_{n-1} \int_{(e_0 + e_1 + \ldots + e_{n-1})} \phi(x)\, dx,$$

and therefore $\int_{(E)} f^{(m)}(x)\, \phi(x)\, dx = AM_m$, where M_m is some number between the greatest and the least of the numbers $\int_{(E_s)} \phi(x)\, dx$; E_s denoting

$$e_0 + e_1 + \ldots + e_{s-1},$$

for $s = 1, 2, \ldots n$. The set E_s is that set for which $A \geq f^{(m)}(x) > a_s$.

The numbers $a_0, a_1, a_2, \ldots a_n$ can be so chosen that the greatest of the differences of the successive numbers is $< \epsilon_m$, where $\{\epsilon_m\}$ is a sequence of decreasing numbers converging to zero. We have now

---

* See Lebesgue, *Annales de Toulouse* (3), vol. I (1909), p. 37.

$$\left| \int_{(E)} f^{(m)}(x)\, \phi(x)\, dx - \int_{(E)} f(x)\, \phi(x)\, dx \right| < \epsilon_m \int_{(E)} |\phi(x)|\, dx,$$

which converges to zero, as $m \sim \infty$. The numbers AM_m consequently converge to $\int_{(E)} f(x)\, \phi(x)\, dx$, as $m \sim \infty$: and thus $\int_{(E)} f(x)\, \phi(x)\, dx = AM$, where M is between the greatest and the least of the numbers $\int_{(e)} \phi(x)\, dx$, and e belongs to a family of sets such that in any one of them $f(x)$ exceeds a certain number between A and zero. By the theorem of § 417, there exists a set E_1 of the family, for which $\int_{(E_1)} \phi(x)\, dx = M$.

The following theorem* has thus been established:

If E be a set of points which is measurable and bounded, and in any number of dimensions, $\phi(x)$, $f(x)$ are summable in E, $f(x)$ being bounded therein, and ≥ 0, then

$$\int_{(E)} f(x)\, \phi(x)\, dx = A \int_{(E_1)} \phi(x)\, dx,$$

where A is any assigned number $\geq$ the upper boundary of $f(x)$ in E, and E_1 is a part of E dependent on the value of A, and belonging to a family of sets of increasing measure, such that for any one of them $f(x)$ is greater than some fixed number.

This theorem may be regarded as the complete generalization of Bonnet's theorem, for any number of dimensions, and for integration through any measurable and bounded set of points.

If $f(x)$ be no longer restricted to be ≥ 0, we may apply the above theorem to the function $f(x) - B$, where B is any number $\leq L$, the lower boundary of $f(x)$ in E.

The theorem then takes the form:

$$\int_{(E)} f(x)\, \phi(x)\, dx = A' \int_{(E_1)} \phi(x)\, dx + B' \int_{(E-E_1)} \phi(x)\, dx,$$

where A' and B' are such that $A' \geq U$, $B' \leq L$; U, L being the upper, and the lower, boundary of $f(x)$ in E, and E_1 is a part of E such that $f(x)$ is therein greater than some number dependent on A' and B'.

If we take the set E to consist of the points of a linear interval (a, b), and $f(x)$ to be monotone, non-increasing, and ≥ 0, the set E_1 will be the set of points of some interval (a, ξ), and thus $M = \int_a^\xi \phi(x)\, dx$, for some value of ξ in (a, b). Similarly, if $f(x)$ be ≥ 0, and monotone, non-decreasing, we must take for E_1 some interval (ξ', b). Thus the first theorem given above reduces to Bonnet's theorem.

---

* See Lebesgue, *Annales sc. de l'école normale* (3), vol. XXVII (1910), p. 443.

426. It will now be shewn that, in case* the set E is the set of points in the plane rectangle $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$, and when the function $f(x^{(1)}, x^{(2)})$ is quasi-monotone and non-negative, of any of the four types considered in § 255, the set E_1, in § 425, may be replaced by the set of points in a plane rectangle with one corner at one of the corners of $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$.

Let $f(x^{(1)}, x^{(2)})$ be quasi-monotone, non-negative and non-diminishing with respect to $(x^{(1)}, x^{(2)})$, and also with respect to $x^{(1)}$ and $x^{(2)}$.

We have, from the theorem of § 448[1]

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) = \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, dG(x^{(1)}, x^{(2)}),$$

where $G(x^{(1)}, x^{(2)})$ denotes

$$\int_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, x^{(2)})} g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}).$$

and this may be transformed by means of the formula (B_1) of § 448, changing the two functions f, g, there employed, into G, f. The three integrals in this formula are all less than, or equal to, the values which would be obtained by substituting for the integrands the value U of the upper boundary of $\Delta_{(x^{(1)}, x^{(2)})}^{(b^{(1)}, b^{(2)})} G(x^{(1)}, x^{(2)})$, and their values are greater than, or equal to, those which would be obtained by substituting for the integrands the lower boundary L of the same expression. Hence the sum of the three integrals is not greater than $U\{f(b^{(1)}, b^{(2)}) - f(a^{(1)}, a^{(2)})\}$, and not less than $L\{f(b^{(1)}, b^{(2)}) - f(a^{(1)}, a^{(2)})\}$.

Also the term $\quad f(a^{(1)}, a^{(2)})\, \Delta_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} G(x^{(1)}, x^{(2)})$

is $\leq U f(a^{(1)}, a^{(2)})$, and is $\quad L f(a^{(1)}, a^{(2)})$. It now follows that

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, dG(x^{(1)}, x^{(2)})$$

is equal to $\quad M\{f(b^{(1)}, b^{(2)}) - f(a^{(1)}, a^{(2)})\} + Mf(a^{(1)}, a^{(2)})$,

or to $Mf(b^{(1)}, b^{(2)})$, where M is some number in the interval (L, U).

Since $\quad \Delta_{(x^{(1)}, x^{(2)})}^{(b^{(1)}, b^{(2)})} G(x^{(1)}, x^{(2)})$

assumes all values between L and U, it follows that

$$M = \Delta_{(\xi^{(1)}, \xi^{(2)})}^{(b^{(1)}, b^{(2)})} G(x^{(1)}, x^{(2)})$$

for some point $(\xi^{(1)}, \xi^{(2)})$ in the rectangle.

We thus have the formula

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) = f(b^{(1)}, b^{(2)}) \int_{(\xi^{(1)}, \xi^{(2)})}^{(b^{(1)}, b^{(2)})} g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) \quad \ldots\ldots(1),$$

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p 273, where the case of quasi-monotone functions of any number of variables is treated

where $(\xi^{(1)}, \xi^{(2)})$ is some point in the closed rectangle $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. It is clear that in this formula $f(b^{(1)}, b^{(2)})$ may be replaced by any number A which is $\geq f(b^{(1)}, b^{(2)})$, for this only involves a change in value of the function f at the point $(b^{(1)}, b^{(2)})$, the condition that the function is quasi-monotone non-diminishing being conserved.

In case $f(x^{(1)}, x^{(2)})$ is positive, quasi-monotone and (non-diminishing, of one of the other types defined in § 255, we obtain in a similar manner, by employing the formulae (B_2), (B_3), (B_4) of § 448

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$$
$$= f(a^{(1)}, a^{(2)}) \int_{(a^{(1)}, a^{(2)})}^{(\xi^{(1)}, \xi^{(2)})} g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) \quad \ldots\ldots(2),$$

where $f(x^{(1)}, x^{(2)})$ is non-increasing with respect to $x^{(1)}$, and with respect to $x^{(2)}$.

Also, when $f(x^{(1)}, x^{(2)})$ is non-diminishing with respect to $x^{(1)}$, and non-increasing with respect to $x^{(2)}$, we have

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, g(x^{(1)}, x^{(2)})\, d(x^{(1)}. x^{(2)})$$
$$= f(b^{(1)}, a^{(2)}) \int_{(\xi^{(1)}, a^{(2)})}^{(b^{(1)}, \xi^{(2)})} g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) \quad \ldots\ldots(3).$$

When $f(x^{(1)}, x^{(2)})$ is non-increasing with respect to $x^{(1)}$, and non-diminishing with respect to $x^{(2)}$, we have

$$\int_{(a^{(1)}. a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, x^{(2)})\, g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$$
$$= f(a^{(1)}, b^{(2)}) \int_{(a^{(1)}, \xi^{(2)})}^{(\xi^{(1)}, b^{(2)})} g(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) \quad \ldots\ldots(4).$$

In case $f(x^{(1)}, x^{(2)})$ is ≥ 0 but is non-increasing with respect to $(x^{(1)}, x^{(2)})$, the above formulae are applicable, according to the type of the function in regard to monotony with respect to the separate variables $x^{(1)}, x^{(2)}$.

REPEATED LEBESGUE INTEGRALS

427. The L-integral of a function of any number of variables has been defined as the limit of a sequence of simple sums; but it is convenient to represent such an integral as a set of repeated limits of sums with respect to the separate variables, and thus to express the integral as a repeated integral with respect to the separate variables, taken in any order. As an L-integral with a non-negative integrand is the measure of a set of points in space of dimensions one greater than the dimensions of the space in which the function is defined, we first shew that the measure of a set of points, measurable in $p + 1$ dimensions, can be exhibited as the single L-integral of the measure of the p-dimensional section of the set by planes perpendicular to one of the coordinate axes, or more generally as the

q-dimensional L-integral of the measure of a $(p - q + 1)$-dimensional section of the set.

Let E denote a measurable and bounded set of points, in the $(p + 1)$-dimensional space between the two planes $x^{(1)} = a^{(1)}$, $x^{(1)} = b^{(1)}$, perpendicular to the $x^{(1)}$-axis Denote by $E(x^{(1)})$ the p-dimensional set which is the section of E by the plane corresponding to any fixed value of $x^{(1)}$ in the interval $(a^{(1)}, b^{(1)})$.

The set E is contained in a set of closed cells Δ, of measure $< m(E) + \epsilon$, where ϵ is an arbitrarily small positive number (see § 127). Giving to ϵ the values in a sequence of decreasing numbers that converges to zero, we have a sequence $\{\Delta_\iota\}$, of sets of closed cells; and the sequence may be so chosen that each set contains the next. The inner limiting set of $\{\Delta_\iota\}$ is a set E_1 which contains E, and is such that $m(E_1) = m(E)$.

Let $\Delta_{\iota 1}, \Delta_{\iota 2}, \dots \Delta_{\iota n}, \dots$ be the cells of the set Δ_ι; we have then

$$m(\Delta_\iota) = \sum_{n=1}^{\infty} \int_a^b m\{\Delta_{\iota n}(x^{(1)})\}\, dx^{(1)},$$

where $\Delta_{\iota n}(x^{(1)})$ is the section of the cell $\Delta_{\iota n}$ by a plane perpendicular to the $x^{(1)}$-axis, and the measure $m\{\Delta_{\iota n}(x^{(1)})\}$ is the p-dimensional measure of $\Delta_{\iota n}(x^{(1)})$. Since an L-integral is completely additive, we have

$$m(\Delta_\iota) = \int_a^b m\{\Delta_\iota(x^{(1)})\}\, dx^{(1)}.$$

In virtue of the theorem of § 398, we now have

$$m(E) = m(E_1) = \lim_{\iota\sim\infty} m(\Delta_\iota) = \int_a^b m\{E_1(x^{(1)})\}\, dx^{(1)},$$

since $\lim\limits_{\iota\sim\infty} m\{\Delta_\iota(x^{(1)})\} = m\{E_1(x^{(1)})\}$.

Again, the set $E_1 - E$, which is of measure zero, is contained in each set of a sequence of sets of cells $\{\Delta_\iota'\}$, where Δ_ι' can be so determined as to be contained in the set Δ_ι. This sequence of cells has an inner limiting set E_2 which contains $E_1 - E$.

We have now, as before,

$$m(\Delta_\iota') = \int_a^b m\{\Delta_\iota'(x^{(1)})\}\, dx^{(1)},$$

and therefore, taking the limit of both sides of the equation, we have

$$\int_a^b m\{E_2(x^{(1)})\}\, dx^{(1)} = \lim_{\iota\sim\infty} m(\Delta_\iota') = 0.$$

It follows that, for almost all values of $x^{(1)}$, $m\{E_2(x^{(1)})\} = 0$; and since $E_2(x^{(1)})$ contains the section of $E_1 - E$, $m\{E(x^{(1)})\}$ exists, for almost all values of $x^{(1)}$, and is equal to $m\{E_1(x^{(1)})\}$, which exists everywhere, as $E_1(x^{(1)})$ is the limit of a sequence of measurable sets.

We have now

$$m(E) = \int_a^b m\{E_1(x^{(1)})\}\,dx^{(1)} = \int_a^b m\{E(x^{(1)})\}\,dx^{(1)}.$$

The following theorem has now been established:

If E be a bounded measurable set of points in $(p+1)$-dimensional space, its section $E(x)$ is measurable, as a p-dimensional set, for almost all values of x. The function $m\{E(x)\}$, of x, defined for almost all values of x, is linearly measurable and its L-integral with respect to x is equal to $m(E)$. The co-ordinate x may be any one of the $p+1$ coordinates which determine the points of E.

Instead of taking x to be one of the coordinates $x^{(1)}, x^{(2)}, \dots x^{(p+1)}$, it may be taken to be typical of a group $(x^{(1)}, x^{(2)}, \dots x^{(q)})$, of these coordinates, where q has any value $< p+1$. The section $E(x)$ will then denote that $(p+1-q)$-dimensional set which is the section of the set E obtained by giving $x^{(1)}, x^{(2)}, \dots x^{(q)}$ fixed values. This will then entail no essential modification in the above proof of the theorem; and accordingly we obtain the following more general result:

If E be a bounded measurable set of points in $(p+1)$-dimensional space, its section $E(x^{(1)}, x^{(2)}, \dots x^{(q)})$, where $q < p+1$, is measurable as a set in $(p+1-q)$-dimensional space, for almost all points $(x^{(1)}, x^{(2)}, \dots x^{(q)})$. The function $m\{E(x^{(1)}, x^{(2)}, \dots x^{(q)})\}$ is measurable as a q-dimensional set, and its L-integral with respect to $(x^{(1)}, x^{(2)}, \dots x^{(q)})$ is equal to $m(E)$. Instead of $x^{(1)}, x^{(2)}, \dots x^{(q)}$, we may take any q of the $p+1$ coordinates $x^{(1)}, x^{(2)}, \dots x^{(p+1)}$.

428. Let $f(x^{(1)}, x^{(2)}, \dots x^{(p)})$ be any bounded function, defined in the cell $(a^{(1)}, a^{(2)}, \dots a^{(p)}; b^{(1)}, b^{(2)}, \dots b^{(p)})$, and measurable in that cell. We first assume that the function is at all points ≥ 0. The integral of the function may be regarded as the measure of a $(p+1)$-dimensional set, defined in a $(p+1)$-dimensional cell $(a^{(1)}, a^{(2)}, \dots a^{(p+1)}, b^{(1)}, b^{(2)}, \dots b^{(p+1)})$, where $a^{(p+1)}$ is any number less than the lower boundary of the function in the cell, and $b^{(p+1)}$ is any number greater than its upper boundary.

We may suppose the function to be defined to have the value zero at any point of the $(p+1)$-dimensional cell, for which it was not originally defined.

We now have, from the foregoing theorem,

$$\int_{(a^{(1)}, a^{(2)}, \dots a^{(p)})}^{(b^{(1)}, b^{(2)}, \dots b^{(p)})} f(x^{(1)}, x^{(2)}, \dots x^{(p)})\, d(x^{(1)}, x^{(2)}, \dots x^{(p)})$$
$$= \int_{(a^{(1)}, a^{(2)}, \dots a^{(q)})}^{(b^{(1)}, b^{(2)}, \dots b^{(q)})} \left\{ \int_{(a^{(q+1)}, a^{(q+2)}, \dots a^{(p)})}^{(b^{(q+1)}, b^{(q+2)}, \dots b^{(p)})} f(x^{(1)}, x^{(2)}, \dots x^{(p)})\, d(x^{(q+1)}, x^{(q+2)}, \dots x^{(p)}) \right\} d(x^{(1)}, x^{(2)}, \dots x^{(p)}),$$

for every value of q $(< p)$.

If we take $q = 1$, we have

$$\int_{(a)}^{(b)} f(x)\,dx = \int_{a^{(1)}}^{b^{(1)}} \left\{ \int_{(a^{(2)},\, a^{(3)},\, \dots\, a^{(p)})}^{(b^{(2)},\, b^{(3)},\, \dots\, b^{(p)})} f(x^{(1)}, x^{(2)}, \dots x^{(p)})\, d(x^{(2)}, x^{(3)}, \dots x^{(p)}) \right\} dx^{(1)}.$$

By repeated use of this result, we express $\int_{(a)}^{(b)} f(x)\,dx$ as a *repeated integral*, taken successively with respect to each coordinate; and it is clear that the order in which the integrations are taken is immaterial.

In particular, we see that

$$\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) = \int_{a^{(1)}}^{b^{(1)}} \left\{ \int_{a^{(2)}}^{b^{(2)}} f(x^{(1)}, x^{(2)})\, dx^{(2)} \right\} dx^{(1)}$$
$$= \int_{a^{(2)}}^{b^{(2)}} \left\{ \int_{a^{(1)}}^{b^{(1)}} f(x^{(1)}, x^{(2)})\, dx^{(1)} \right\} dx^{(2)}.$$

If the function $f(x^{(1)}, \dots x^{(p)})$ is no longer restricted to be non-negative, it can be expressed as the difference of two measurable non-negative functions, and the general theorem may be applied to each of these functions. We have then the following result:

If $f(x^{(1)}, x^{(2)})$ *be defined in the cell* $(a^{(1)}, a^{(2)}, b^{(1)}, b^{(2)})$, *and be bounded and measurable, the two repeated integrals*

$$\int_{a^{(1)}}^{b^{(1)}} \int_{a^{(2)}}^{b^{(2)}} f(x^{(1)}, x^{(2)})\, dx^{(1)} dx^{(2)}, \qquad \int_{a^{(2)}}^{b^{(2)}} \int_{a^{(1)}}^{b^{(1)}} f(x^{(1)}, x^{(2)})\, dx^{(2)} dx^{(1)},$$

both exist, and are equal to $\int_{(a^{(1)},\, a^{(2)})}^{(b^{(1)},\, b^{(2)})} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$.

The form $\int_{a^{(1)}}^{b^{(1)}} \int_{a^{(2)}}^{b^{(2)}} f(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)}$ is employed to denote the repeated integral $\int_{a^{(1)}}^{b^{(1)}} \left\{ \int_{a^{(2)}}^{b^{(2)}} f(x^{(1)}, x^{(2)})\, dx^{(2)} \right\} dx^{(1)}$, in which the integration is taken, first with respect to $x^{(2)}$, and then with respect to $x^{(1)}$.

It has been shewn that $\int_{a^{(2)}}^{b^{(2)}} f(x^{(1)}, x^{(2)})\, dx^{(2)}$ has a definite value for almost all values of $x^{(1)}$.

More generally we have the theorem that:

The L-integral of a function of p variables, defined in a cell (a, b), *and bounded and measurable in that cell, is equivalent to the repeated integral obtained by integrating the function successively with regard to the p-coordinates, taken in any order.*

429. In order to extend the results of § 428 to the case of an unbounded function, it will be sufficient to consider the case in which the function is non-negative, since any summable function can be treated as the difference of two such functions.

In the two-dimensional case, let $f(x^{(1)}, x^{(2)})$ be a non-negative function, summable over the measurable set E. Let $\{k_n\}$ be a monotone sequence of positive numbers without upper limit, and let $f_n(x^{(1)}, x^{(2)}) = f(x^{(1)}, x^{(2)})$, when $f(x^{(1)}, x^{(2)}) \leq k_n$, and let $f_n(x^{(1)}, x^{(2)}) = k_n$, when $f(x^{(1)}, x^{(2)}) > k_n$.

We have then, from § 428,

$$\int_{(E)} f_n(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}) = \int dx^{(1)} \int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}.$$

If $\int_{(E)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$ exists, it is defined as

$$\lim_{n\sim\infty} \int_{(E)} f_n(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}),$$

which is equal to $\quad \lim_{n\sim\infty} \int dx^{(1)} \int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}.$

Denoting $\int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}$ by $s_n(x^{(1)})$, we see that $\{s_n(x^{(1)})\}$ is a monotone non-decreasing sequence, for each value of $x^{(1)}$; and on the assumption that $\int_{(E)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)})$ exists, we conclude that $\lim_{n\sim\infty} \int s_n(x^{(1)})\, dx^{(1)}$ exists, and has the same value as the integral.

Applying the theorem of § 399, we see that $\{s_n(x^{(1)})\}$ converges for almost all points $x^{(1)}$, and also that

$$\int s(x^{(1)})\, dx^{(1)} = \int_{(E)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}):$$

where $\quad s(x^{(1)}) = \lim_{n\sim\infty} s_n(x^{(1)}) = \lim_{n\sim\infty} \int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}.$

Applying again the theorem of § 399, we see that, for each value of $x^{(1)}$ for which $\lim_{n\sim\infty} \int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}$ is finite,

$$\int f(x^{(1)}, x^{(2)})\, dx^{(2)} = \lim_{n\sim\infty} \int f_n(x^{(1)}, x^{(2)})\, dx^{(2)}.$$

Therefore $\quad \int dx^{(1)} \int f(x^{(1)}, x^{(2)})\, dx^{(2)}$

exists, and is equal to

$$\int_{(E)} f(x^{(1)}, x^{(2)})\, d(x^{(1)}, x^{(2)}).$$

The complete result may be stated as follows:

If $f(x^{(1)}, x^{(2)})$ be any function, bounded or unbounded, that is integrable (L) over the measurable and bounded set E, then the repeated integrals

$$\int dx^{(1)} \int f(x^{(1)}, x^{(2)})\, dx^{(2)}, \quad \int dx^{(2)} \int f(x^{(1)}, x^{(2)})\, dx^{(1)}$$

both exist, and have the same value as the L-integral function over E.

It is clear that this statement includes the case of a function of any number of variables, as $x^{(1)}$ and $x^{(2)}$ may each be regarded as typical of a set of variables. The first proof of this result was given by Fubini*, who deduced the theorem from a definition due to Lebesgue, of the superior and inferior integrals of a function, which are not identical with Darboux's upper and lower integrals, defined in § 331. The above proof of the theorem† was given by Hobson, who also gave the following extension:

If $f(x^{(1)}, x^{(2)})$ *be measurable in the measurable and bounded set* E, *and if*

$$\int dx^{(1)} \int |f(x^{(1)}, x^{(2)})| \, dx^{(2)}$$

has a definite value, then $\int_{(E)} f(x^{(1)}, x^{(2)}) \, d(x^{(1)}, x^{(2)})$ *exists as an L-integral, and is consequently equal to the repeated integrals of* $f(x^{(1)}, x^{(2)})$.

Let us consider the function $\int |f(x^{(1)}, x^{(2)})| \, dx^{(2)}$ and apply the last part of the theorem of § 399 Thus, if $\int s(x^{(1)}) \, dx^{(1)}$ exists, it follows that $\lim\limits_{n\sim\infty} \int s_n(x^{(1)}) \, dx^{(1)}$ exists, and is equal to $\int s(x^{(1)}) \, dx^{(1)}$.

Also

$$\int s_n(x^{(1)}) \, dx^{(1)} = \int dx^{(1)} \int |f_n(x^{(1)}, x^{(2)})| \, dx^{(2)} = \int_{(E)} |f_n(x^{(1)}, x^{(2)})| \, d(x^{(1)}, x^{(2)});$$

where $s_n(x^{(1)})$ denotes $\int |f_n(x^{(1)}, x^{(2)})| \, dx^{(2)}$.

If we assume that $\int dx^{(1)} \int |f(x^{(1)}, x^{(2)})| \, dx^{(2)}$ exists, we see that

$$\int_{(E)} |f(x^{(1)}, x^{(2)})| \, d(x^{(1)}, x^{(2)}),$$

which is defined as

$$\lim_{n\sim\infty} \int_{(E)} f_n(x^{(1)}, x^{(2)}) \, d(x^{(1)}, x^{(2)}), \text{ or as } \lim_{n\sim\infty} \int s_n(x^{(1)}) \, dx^{(1)},$$

exists, and is equal to $\int s(x^{(1)}) \, dx^{(1)}$, or $\int dx^{(1)} \int |f(x^{(1)}, x^{(2)})| \, dx^{(2)}$.

It then follows that $f(x^{(1)}, x^{(2)})$ is summable in E.

In the case in which $f(x^{(1)}, x^{(2)})$ is bounded in E, the theorem is an immediate corollary from the preceding theorem.

It may happen that the repeated integrals of $f(x^{(1)}, x^{(2)})$ exist, but that $\int dx^{(1)} \int |f(x^{(1)}, x^{(2)})| \, dx^{(2)}$ does not exist. In that case the L-integral of $f(x^{(1)}, x^{(2)})$ does not exist, and the two repeated integrals of $f(x^{(1)}, x^{(2)})$ may have unequal values.

---

* *Rend. del. Real. Accad. dei Lincei* (5), vol. XVI (1910), p. 608.

† *Proc. Lond. Math. Soc.* (2), vol. VIII (1909), p. 22.

It has been proved* by Sierpiński that the existence, for a function $f(x^{(1)}, x^{(2)})$, bounded in the square (0, 0; 1, 1), of the integrals

$$\int_0^1 f(x^{(1)}, x^{(2)})\, dx^{(1)}, \text{ for } 0 \leqq x^{(2)} \leqq 1,$$

and

$$\int_0^1 f(x^{(1)}, x^{(2)})\, dx^{(2)}, \text{ for } 0 \leqq x^{(1)} \leqq 1,$$

does not necessarily entail the existence of the repeated integral

$$\int_0^1 dx^{(1)} \int_0^1 f(x^{(1)}, x^{(2)})\, dx^{(2)};$$

but his proof involves the assumption of the truth of Cantor's hypothesis that $2^{\aleph_0} = \aleph_1$. The hypothesis is not required in the case of an unbounded function.

EXAMPLES

1. The repeated integrals

$$\int_0^1 \int_0^1 \frac{(x^2 - y^2)}{(x^2 + y^2)^2}\, dx\, dy, \quad \int_0^1 \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2}\, dy\, dx$$

both exist, but have unequal values, the function $\dfrac{x^2 - y^2}{(x^2 + y^2)^2}$ not being summable in the square (0, 0; 1, 1).

2. Let $f(x, y) = \dfrac{1}{(x - \frac{1}{2})^3}$, for $0 < y < |x - \frac{1}{2}|$; and let $f(x, y) = 0$, for $y \geqq |x - \frac{1}{2}|$, and for $y = 0$. The integrals $\int_0^1 f(x, y)\, dx$, $\int_0^1 f(x, y)\, dy$ exist as R-integrals and therefore as L-integrals, also $\int_0^1 dy \int_0^1 f(x, y)\, dx$ exists, but $\int_0^1 dx \int_0^1 f(x, y)\, dy$ does not exist. This example was given by Sierpiński.

A FUNDAMENTAL APPROXIMATION THEOREM

430. The following theorem affords the means of relating the L-integral of a summable function with a sequence of integrals of continuous functions, in such a way that various properties of the L-integral may be obtained by considering the special case in which the integrand is a continuous function:

If† $f(x)$ *be summable in a given interval, or cell,* (a, b), *a continuous function* $\phi(x)$ *can be constructed which satisfies the condition that*

$$\int_a^b |f(x) - \phi(x)|\, dx$$

is less than an arbitrarily prescribed positive number. Moreover, in case $f(x) \geqq 0$, *in* (a, b), *the function* $\phi(x)$ *can be so determined that* $\phi(x) \geqq 0$, *in* (a, b).

The proof of this theorem may be exhibited as a process which has several stages.

---

* *Fundamenta Math.* vol. I (1920), p. 142.

† See Hobson, *Proc. Lond. Math. Soc.* (2), vol. XII (1913), p. 156.

(1). Consider the case in which $f(x) = 1$, in the cell, or interval, (α, β) contained in (a, b), and is elsewhere zero.

Let the continuous function $\phi(x)$ be defined to be equal to 1 in the cell

$$(\alpha^{(1)}, \alpha^{(2)}, \dots; \beta^{(1)}, \beta^{(2)}, \dots),$$

and to be equal to $1 - \theta$ on the boundary of the cell

$$(\alpha^{(1)} - \theta\epsilon, \alpha^{(2)} - \theta\epsilon, \dots; \beta^{(1)} + \theta\epsilon, \beta^{(2)} + \theta\epsilon, \dots),$$

for each value of θ such that $0 \leq \theta \leq 1$; and outside the cell for which $\theta = 1$, let $\phi(x) = 0$ We may disregard all points that are not in the fundamental cell (a, b).

The function $\phi(x)$ has values between 0 and 1 in the set of points between the two cells (α, β) and $(\alpha - \epsilon, \beta + \epsilon)$; and it is continuous in (a, b). Since $f(x) - \phi(x)$ vanishes in the cell (α, β), we see that

$$\int_a^b | f(x) - \phi(x) | \, dx$$

is less than $A\epsilon$, where A is a fixed number independent of ϵ; and this is arbitrarily small.

(2) Let $f(x) = 1$, in each cell of a finite set $\Delta_1, \Delta_2, \dots \Delta_n$, of non-overlapping cells contained in (a, b), and let $f(x) = 0$ at all other points of (a, b). Let $\phi_r(x)$ be the continuous function, defined as in (1), such that, if $f_r(x) = 1$ in Δ_r, and everywhere else $= 0$,

$$\int_a^b | f_r(x) - \phi_r(x) | \, dx < \frac{1}{n} \epsilon;$$

also suppose $\phi_r(x)$ defined for $r = 1, 2, 3, \dots n$. Let

$$\phi(x) = \phi_1(x) + \phi_2(x) + \dots + \phi_n(x),$$

we have then

$$\int_a^b | f(x) - \phi(x) | \, dx < \sum_{r=1}^{r=n} \int_a^b | f_r(x) - \phi_r(x) | \, dx < \epsilon.$$

The function $\phi(x)$ is continuous, and satisfies the condition in the theorem for the particular function $f(x)$.

(3). Let $f(x) = 1$ in a measurable set of points e, contained in (a, b); and let $f(x) = 0$ in $C(e)$.

A set of non-overlapping cells $\Delta_1, \Delta_2, \dots \Delta_m, \dots$, of total measure exceeding $m(e)$ by less than $\frac{1}{4}\epsilon$, can be so determined that they contain all the points of e either within them or on their boundaries. An integer n can be so chosen that $\sum_{r=n+1}^{\infty} m(\Delta_r) < \frac{1}{4}\epsilon$, and that

$$0 < \sum_{r=1}^{r=n} m(\Delta_r) - m(e) < \frac{1}{4}\epsilon.$$

Let $f_n(x)$ be the function that has the value 1 in each of the cells

$$\Delta_1, \Delta_2, \dots \Delta_n,$$

and elsewhere the value zero. In virtue of (2) a continuous function $\phi(x)$ can be so determined that

$$\int_a^b |f_n(x) - \phi(x)|\, dx < \tfrac{1}{2}\epsilon.$$

We have also $$\int_a^b |f(x) - f_n(x)|\, dx < \tfrac{1}{2}\epsilon,$$

and therefore $$\int_a^b |f(x) - \phi(x)|\, dx < \epsilon.$$

The continuous function $\phi(x)$ satisfies the condition of the theorem for the particular function $f(x)$.

(4). Let $e_1, e_2, \ldots e_m$ be m measurable sets, no two of which have a point in common, and all of which are in the cell (a, b). Let numbers $c_1, c_2, \ldots c_m$ be prescribed, and let $f(x)$ be the function that has the value c_r in the set e_r, for the values $1, 2, 3, \ldots m$, of r, and has the value zero at points that do not belong to any of the sets e_r.

Let $\phi_r(x)$ be the continuous function determined as in (3), such that, if $f_r(x) = 1$, in e_r, and is elsewhere zero,

$$\int |f_r(x) - \phi_r(x)|\, dx < \frac{\epsilon}{|c_1| + |c_2| + \ldots + |c_m|};$$

and let these functions $\phi_r(x)$ be constructed for the values $r = 1, 2, 3, \ldots m$.

If $\phi(x) = c_1\phi_1(x) + c_2\phi_2(x) + \ldots + c_m\phi_m(x)$, we have

$$\int_a^b |f(x) - \phi(x)|\, dx \leq \sum_{r=1}^{r=m} |c_r| \int_a^b |f_r(x) - \phi_r(x)|\, dx < \epsilon.$$

Therefore $\phi(x)$ is the continuous function required by the theorem, which corresponds to the function $f(x)$.

(5). Let $f(x)$ be a bounded and summable function. In accordance with § 385, a function $\bar{f}(x)$ can be determined which is such that

$$|f(x) - \bar{f}(x)| < \eta,$$

and that $\bar{f}(x)$ has only a finite number of distinct values, each of which it takes in a measurable set of points. Let $\phi(x)$ be the continuous function, constructed as in (4), such that

$$\int_a^b |\bar{f}(x) - \phi(x)|\, dx < \epsilon.$$

We now have

$$\int_a^b |f(x) - \phi(x)|\, dx \leq \int_a^b |f(x) - \bar{f}(x)|\, dx + \int_a^b |\bar{f}(x) - \phi(x)|\, dx$$
$$< \eta m(a, b) + \epsilon;$$

where $m(a, b)$ is the measure of the fundamental cell (a, b).

Since η and ϵ are arbitrarily small, a continuous function $\phi(x)$ such as the theorem requires can be determined, corresponding to any bounded function that is summable in (a, b).

(6). Lastly, let $f(x)$ be unbounded, but summable in (a, b). A bounded summable function $f_N(x)$ can be determined, as in § 386, such that

$$\int_a^b |f(x) - f_N(x)|\, dx$$

is arbitrarily small, say $< \frac{1}{2}\epsilon$; then if $\phi(x)$ be a continuous function, obtained as in (5), such that

$$\int_a^b |f_N(x) - \phi(x)|\, dx < \tfrac{1}{2}\epsilon,$$

we have
$$\int_a^b |f(x) - \phi(x)|\, dx < \epsilon.$$

This completes the proof of the theorem, for the general case of any summable function $f(x)$ It is clear, by examining the successive stages of the process of construction of $\phi(x)$, that if $f(x) \geq 0$, in (a, b), then $\phi(x) \geq 0$, in (a, b).

We observe that

$$\left|\int_a^b \{f(x) - \phi(x)\}\, dx\right| \leq \int_a^b |f(x) - \phi(x)|\, dx < \epsilon.$$

If ϵ have successively the values in a diminishing sequence $\{\epsilon_r\}$ that converges to zero, and if $\phi_r(x)$ be the continuous function that corresponds to the value ϵ_r, of ϵ, we see that

$$\lim_{r\sim\infty} \int_a^b \phi_r(x)\, dx = \int_a^b f(x)\, dx,$$

and also
$$\lim_{r\sim\infty} \int_a^b |f(x) - \phi_r(x)|\, dx = 0.$$

The following deduction from the approximation theorem is of importance in view of applications:

If $f(x)$ be summable in a given interval, or cell, (a, b), a function $\psi(x)$ which has a constant value within each interval, or cell, of some finite set into which (a, b) is divided, can be so determined that $\int_a^b |f(x) - \psi(x)|\, dx$ is less than an arbitrarily prescribed positive number.

On the boundaries of the cells of the finite set, $f(x)$ may have arbitrarily assigned values, as these do not affect the value of the integral.

Taking, in accordance with the above theorem, the continuous function $\phi(x)$, such that $\int_a^b |f(x) - \phi(x)| < \frac{1}{2}\epsilon$, we may divide (a, b) into a finite set of intervals, or cells, such that, in each of them, the fluctuation of $\phi(x)$ is $< \dfrac{\epsilon}{2m(a, b)}$ (see § 217).

Let $\psi(x)$ have, within each such interval, or cell, the value of $\phi(x)$ at its centre; then $\int_a^b |\phi(x) - \psi(x)|\, dx < \frac{1}{2}\epsilon$.

It now follows that $\int_a^b |f(x) - \psi(x)|\,dx < \epsilon$, and thus $\psi(x)$ satisfies the required condition.

A sequence $\{f_r(x)\}$ of functions of the type of $\psi(x)$ exists, such that

$$\lim_{r\sim\infty}\int_a^b |f(x) - \psi_r(x)|\,dx = 0.$$

431. Various theorems of importance in the theories of integration and of series can be deduced from the theorem of § 430.

Let $f(x)$ be summable in the open linear interval (α, β), and let (a, b) be a closed interval contained in (α, β). Consider

$$\int_a^b |f(x+t) - f(x)|\,dx,$$

where t is $< \beta - b$, so that the integrand is defined in (a, b). Let $\phi(x)$ be the continuous function, determined as in the theorem of § 430, so that

$$\int_\alpha^\beta |f(x) - \phi(x)|\,dx < \tfrac{1}{3}\epsilon.$$

The integral of $|f(x+t) - f(x)|$ over (a, b) does not exceed the sum of the integrals of

$$|f(x+t) - \phi(x+t)|,\quad |\phi(x+t) - \phi(x)|,\quad \text{and}\quad |\phi(x) - f(x)|$$

over the same interval. The first and the third of these integrals are each $< \frac{1}{3}\epsilon$, and the second integral is also $< \frac{1}{3}\epsilon$, provided t is less than a fixed number δ_ϵ, dependent on ϵ. Therefore

$$\int_a^b |f(x+t) - f(x)|\,dx < \epsilon,$$

if $t < \delta_\epsilon$; and since ϵ is arbitrary, it follows that the limit of the value of the integral, as $t \sim 0$, is zero.

We have thus obtained the following theorem, which has been established in a different manner* by Lebesgue:

If $f(x)$ be summable in an open linear interval (α, β), and if (a, b) be a closed interval interior to (α, β), then

$$\lim_{t\sim 0}\int_a^b |f(x+t) - f(x)|\,dx = 0.$$

The corresponding theorem holds for a summable function defined in an open cell (α, β) of any number of dimensions; thus, in the case of a function $f(x^{(1)}, x^{(2)})$ of two dimensions, we have

$$\lim_{t^{(1)}\sim 0,\, t^{(2)}\sim 0}\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} |f(x^{(1)} + t^{(1)}, x^{(2)} + t^{(2)}) - f(x^{(1)}, x^{(2)})|\, d(x^{(1)}, x^{(2)}) = 0.$$

No essential change in the proof of the theorem for the linear case is required to make this extension.

---

* See the *Leçons sur les séries trigonométriques*, p. 15

432. The following theorem, also due to Lebesgue*, may be deduced from the theorem of § 430:

If $f(x)$ be summable in a linear interval (a, b), then, for almost all points x_0, in (a, b), $|f(x)-\alpha|$ is, for $x = x_0$, the differential coefficient of its indefinite integral, whatever value the constant α may have. In particular

$$\int_{x_0}^{x} |f(x)-f(x_0)|\, dx$$

has a differential coefficient at $x = x_0$, equal to zero, for almost all values of x_0, in (a, b).

The integral
$$\int_{x_0}^{x_0+h} |f(x)-\alpha|\, dx$$
is between the two numbers
$$\int_{x_0}^{x_0+h} |\phi(x)-\alpha|\, dx \pm \int_{x_0}^{x_0+h} |f(x)-\phi(x)|\, dx,$$
where $\phi(x)$ is a continuous function which we may take to be such that
$$\int_a^b |f(x)-\phi(x)|\, dx < \epsilon^2.$$

It follows from this inequality that the set of points at which
$$|f(x)-\phi(x)| \geq \epsilon,$$
has its measure less than ϵ, and therefore, if x_0 be any point of a certain set H_ϵ, where $m(H_\epsilon) > b-a-\epsilon$, the condition $|f(x_0)-\phi(x_0)| < \epsilon$ is satisfied.

Since $|\phi(x)-\alpha|$ is a continuous function,
$$\lim_{h\sim 0} \frac{1}{h}\int_{x_0}^{x_0+h} |\phi(x)-\alpha|\, dx$$
exists, and is equal to $|\phi(x_0)-\alpha|$, for all points x_0 interior to (a, b).

For almost all values of x_0,
$$\lim_{h\sim 0} \frac{1}{h}\int_{x_0}^{x_0+h} |f(x)-\phi(x)|\, dx$$
exists, and is equal to $|f(x_0)-\phi(x_0)|$, since $|f(x)-\phi(x)|$ is a summable function (see § 405).

We now see that
$$\overline{\underline{\lim}}_{h\sim 0} \frac{1}{h}\int_{x_0}^{x_0+h} |f(x)-\alpha|\, dx$$
is between the numbers
$$|\phi(x_0)-\alpha| \pm |f(x_0)-\phi(x_0)|,$$
for almost all values of x_0, whatever value α may have. In a set K_ϵ, such that
$$m(K_\epsilon) = m(H_\epsilon) > b-a-\epsilon,$$

---

* *Loc. cit.* p. 13.

it follows that $$\overline{\lim_{h \to 0}} \frac{1}{h} \int_{x_0}^{x_0+h} |f(x) - \alpha| \, dx$$

is between the numbers $$|f(x_0) - \alpha| \pm 2\epsilon.$$

Let ϵ have the values in a sequence which converges to zero, then $m(K_\epsilon)$ converges to $b - a$, we see then that, for almost all values of x_0,

$$\lim_{h \sim 0} \frac{1}{h} \int_{x_0}^{x_0+h} |f(x) - \alpha| \, dx$$

exists, and is equal to $$|f(x_0) - \alpha|,$$

whatever value α may have. The theorem has now been established.

433 Let $f_1(x)$ be a function which is summable, and is ≥ 0, in a cell, or interval, (a, b); let us further assume that $\{f_1(x)\}^2$ is summable in (a, b). A continuous function $\phi_1(x)$, at every point ≥ 0, can be so determined that

$$\int_a^b |\{f_1(x)\}^2 - \{\phi_1(x)\}^2| \, dx$$

is less than $\epsilon/4$. We have, for every value of x,

$$\{f_1(x) - \phi_1(x)\}^2 \leq |\{f_1(x)\}^2 - \{\phi_1(x)\}^2|;$$

and it follows that $$\int_a^b \{f_1(x) - \phi_1(x)\}^2 \, dx < \epsilon/4.$$

Next, let $f(x)$ be no longer restricted to be ≥ 0, in (a, b), but let $\{f(x)\}^2$ be summable in that cell, or interval; $f(x)$ can then be expressed as the difference $f_1(x) - f_2(x)$ of two functions, each of which is ≥ 0, in (a, b), and such that the square of each of them is summable in (a, b).

Let $\phi_2(x)$ be a continuous function, corresponding to $f_2(x)$, such that

$$\int_a^b \{f_2(x) - \phi_2(x)\}^2 \, dx < \epsilon/4;$$

and let $$\phi(x) = \phi_1(x) - \phi_2(x).$$

We have now

$$\int_a^b \{f(x) - \phi(x)\}^2 \, dx \leq 2 \int_a^b \{f_1(x) - \phi_1(x)\}^2 \, dx + 2 \int_a^b \{f_2(x) - \phi_2(x)\}^2 \, dx,$$

and the expression on the right-hand side is $< \epsilon$.

The following theorem has thus been established:

If $\{f(x)\}^2$ be summable (whether it be bounded or unbounded) in the interval, or cell, (a, b), a continuous function $\phi(x)$ can be so determined that

$$\int_a^b \{f(x) - \phi(x)\}^2 \, dx$$

is less than a prescribed positive number ϵ.

As in § 430, it can be shewn that the continuous function $\phi(x)$ may be replaced by a function $\psi(x)$ which has a constant value in each interval, or cell, of a finite set into which (a, b) may be divided.

From this theorem a result can be deduced which has been otherwise obtained, for the case of a linear interval, by A. C. Dixon*.

If $\{f(x)\}^2$ *be summable, in an open cell, or interval,* (α, β), *and* (a, b) *be a cell, or an interval, contained in* (α, β), *then*

$$\lim_{x\sim 0}\int_a^b \{f(t+x) - f(t)\}^2\, dt = 0.$$

In the case of a cell of two dimensions, x *represents a pair* $(x^{(1)}, x^{(2)})$ *of numbers, and the limit is taken as* $x^{(1)}, x^{(2)}$ *converge independently of one another to zero.*

We assume x to be such that $t + x$ is in (α, β), for all values of t in (a, b). If $\phi(t)$ be a continuous function such that

$$\int_\alpha^\beta \{f(t) - \phi(t)\}^2\, dt < \epsilon,$$

we have
$$\int_a^b \{f(t+x) - f(t)\}^2\, dt < 3\int_a^b \{f(t+x) - \phi(t+x)\}^2\, dt$$
$$+ 3\int_a^b \{\phi(t+x) - \phi(t)\}^2\, dt$$
$$+ 3\int_a^b \{\phi(t) - f(t)\}^2$$

The first and third integrals on the right-hand side are each $< \epsilon$ The function $\phi(t)$ being uniformly continuous in a closed interval (a', b') which contains (a, b), and is in (α, β), we see that if $|x|$, or in the two-dimensional case, $|x^{(1)}|$ and $|x^{(2)}|$, be sufficiently small, the second integral is $< \epsilon$. It follows that

$$\int_a^b \{f(x+t) - f(t)\}^2\, dt < 9\epsilon,$$

if $|x|$ be sufficiently small. Since ϵ is arbitrary, the theorem has been established.

From this result, the following theorem† may be deduced:

If $\{f(x)\}^2$ *and* $\{\phi(x)\}^2$ *be summable in an open interval, or cell,* (α, β), *and* (a, b) *be contained in* (α, β),

$$\int_a^b \phi(t)\, f(t+x)\, dt$$

is a continuous function of x, *for all points* x *in* (a, b).

In the two-dimensional case, $f(t+x)$ *denotes* $f(t^{(1)} + x^{(1)}, t^{(2)} + x^{(2)})$.

---

* *Proc. Camb. Phil. Soc.* vol. xv (1910), p. 210.

† For the case of a linear interval, this theorem has been proved in a different manner by A. C. Dixon, *loc. cit.*, in the case in which the functions ϕ, f are identical.

Let $F(x)$ denote $\int_a^b \phi(t) f(t+x)\, dt$;

then, employing Schwarz's inequality, we have

$$\{F(x_1) - F(x_2)\}^2 \leq \int_a^b \{\phi(t)\}^2 dt \int_a^b \{f(t+x_1) - f(t+x_2)\}^2 dt;$$

the second integral on the right-hand side is less than

$$\int_{\alpha_1}^{\beta_1} \{f(t+x') - f(t)\}^2 dt,$$

if (α_1, β_1) is an interval or cell contained in (α, β), and x' denotes $x_1 - x_2$. This integral converges to zero, with $x_1 - x_2$; hence $|F(x_1) - F(x_2)|$ is arbitrarily small, for all sufficiently small values of $|x_1 - x_2|$, and therefore $F(x)$ is a continuous function.

APPROXIMATE REPRESENTATION OF AN L-INTEGRAL AS A RIEMANN SUM

434. Let $f(x)$ be summable in the interval, or cell, (a, b), and let $f_N(x)$ denote the function that has the same values as $f(x)$ at all points at which $|f(x)| \leq N$, and which has the value N, or $-N$, at all other points of (a, b). Let the measure of the set E, of points at which $|f(x)| > N$, have the value ϵ. By the theorem of § 430, a continuous function $\phi(x)$ can be so determined that $\int_a^b |f_N(x) - \phi(x)|\, dx < \epsilon^2$. In a certain set of points L, such that $m(L) < \epsilon$, we may have $|f_N(x) - \phi(x)| \geq \epsilon$, in the complementary set $C(L)$ we have $|f_N(x) - \phi(x)| < \epsilon$. The function $\phi(x)$ may be taken to be such that $|\phi(x)| \leq N$; for if it have values numerically $> N$, these may be replaced by N or $-N$, without affecting the conditions that $\phi(x)$ is to satisfy.

A net can be fitted on to the interval, or cell, (a, b), such that

$$\sum_{r=1}^{r=m} \phi(\xi_r)\, m(\delta_r)$$

differs from $\int_a^b \phi(x)\, dx$ by less than ϵ; where $\delta_1, \delta_2, \ldots \delta_m$ denote the m meshes of the net, and the points $\{\xi_r\}$ are chosen in any manner so that ξ_r is a point of δ_r. If there are meshes of the net such that all their points belong either to L or to E, the total measure of such meshes cannot exceed 2ϵ.

In all the other meshes there are points that belong neither to L nor to E. In these latter, the points ξ_r can be so chosen that they do not belong to L, or to E. In a mesh δ_r, all the points of which belong to L or to E, we can choose ξ_r so that $|f(\xi_r)|$ differs from the lower boundary of $|f(x)|$ in δ_r, by less than ϵ. Let this be done for all such meshes; then, denoting the set of all such meshes by Δ, where $m(\Delta) \leq 2\epsilon$, we see that

$$|\Sigma f(\xi_r)\, m(\delta_r)| < \epsilon m(\Delta) + \int_{(\Delta)} |f(x)|\, dx.$$

The sum on the left-hand side is here taken for all the meshes of the set Δ. We have also $|\Sigma\phi(\xi_r)\, m(\delta_r)| < 2U\epsilon$; where U is the upper boundary of $|\phi(x)|$ in (a, b).

In the other meshes of the net we have

$$|\Sigma\{f(\xi_r) - \phi(\xi_r)\}\, m(\delta_r)| < \epsilon l:$$

where l is the measure of the interval, or cell, (a, b).

Also $\int_a^b f(x)\, dx$ differs from $\int_a^b \phi(x)\, dx$ by less than $\epsilon^2 + \int_{(E)} |f(x)|\, dx$: hence $\int_a^b f(x)\, dx$ differs from $\sum_{r=1}^{r=m} \phi(\xi_r)\, m(\delta_r)$ by less than

$$\epsilon l + \epsilon^2 + \int_{(E)} |f(x)|\, dx$$

Now $\sum_{r=1}^{r=m} \phi(\xi_r)\, m(\delta_r)$ differs from $\sum_{r=1}^{r=m} f(\xi_r)\, m(\delta_r)$ by less than

$$l\epsilon + 2\epsilon^2 + 2U\epsilon + \int_{(\Delta)} |f(x)|\, dx$$

As $N \sim \infty$, we have $\epsilon \sim 0$, and then $\int_{(\Delta)} |f(x)|\, dx$ and $\int_{(E)} |f(x)|\, dx$ converge to zero. Also $2U\epsilon$ ($= 2N\epsilon$) converges to zero (see § 388).

We now see that $\int_a^b f(x)\, dx$ differs from the Riemann sum

$$\sum_{r=1}^{r=m} f(\xi_r)\, m(\delta_r)$$

by an amount which can be made arbitrarily small by choosing N sufficiently large.

We have thus proved the following theorem, of which Lebesgue* has given another proof:

If $f(x)$ be any function that is summable in the interval, or cell, (a, b), a net can be fitted on to (a, b) such that $\int_a^b f(x)\, dx$ differs arbitrarily little from the Riemann sum $\sum_{r=1}^{r=m} f(\xi_r)\, m(\delta_r)$, provided the points ξ_r be chosen properly in the meshes δ_r of the net.

The chief interest of this theorem arises from the fact that, if we have a finite set of functions $f^{(1)}(x), f^{(2)}(x), \dots f^{(s)}(x)$, all summable in the interval, or cell, (a, b), it is possible to determine a net, and a definite set of points $\xi_1, \xi_2, \dots \xi_m$ in the meshes $\delta_1, \delta_2, \dots \delta_m$, respectively, so that, for each of the s functions, the difference of the integral and the corresponding Riemann sum is less than an arbitrarily chosen number, the points ξ being the same for all the s functions.

---

[*] See *Annales de Toulouse* (3), vol I (1909), p. 33.

To shew that this is the case, only a slight modification in the foregoing proof is required. We have to deal with s sets $E_1, E_2, \dots E_s$, corresponding to a fixed value of N; the number ϵ can be taken to be the greatest of the numbers $m(E_1), m(E_2), \dots m(E_s)$. The meshes of the net which are such that every point of one of them belongs to one at least of the $2s$ sets $m(E_1)$, $m(E_2)$, … $m(E_s)$, $m(L_1)$, $m(L_2)$, … $m(L_s)$ have a total measure $\leq 2s\epsilon$, and these meshes must be considered separately from the others, as in the case $s = 1$; they form a set Δ, such that $m(\Delta) \leq 2s\epsilon$. In one of these meshes δ_r, we take ξ_r so that $\sum_{p=1}^{p=s} |f^{(p)}(\xi_r)|$ differs from the lower boundary of $\sum_{p=1}^{p=s} |f^{(p)}(x)|$ in δ_r, by less than ϵ; we then have, for the set of meshes Δ,

$$|\Sigma f^{(p)}(\xi_r)\, m(\delta_r)| < \epsilon m(\Delta) + \sum_{p=1}^{p=s} \int_{(\Delta)} |f^{(p)}(x)|\, dx$$

for each of the values $p = 1, 2, \dots s$, of p. The preceding proof may then be applied to each of the functions $f^{(1)}(x), \dots f^{(s)}(x)$, without essential change.

We have now obtained the following theorem:

If $f^{(1)}(x), f^{(2)}(x), \dots f^{(s)}(x)$ be functions that are summable in the interval, or cell, (a, b), then a net can be fitted on to (a, b), and a set of points $\xi_1, \xi_2, \dots \xi_m$ can be determined in the m meshes of the net, so that, for each of the s functions, $\int_a^b f(x)\, dx$ differs from $\sum_{r=1}^{r=m} f(\xi_r)\, m(\delta_r)$ by less than an arbitrarily chosen number.

435. By means of this theorem, various algebraical inequalities can be extended so as to obtain corresponding inequalities which involve L-integrals. The simplest case we shall consider is that of the well-known inequality

$$(a_1 b_1 + a_2 b_2 + \dots + a_m b_m)^2 \leq (a_1^2 + a_2^2 + \dots + a_m^2)(b_1^2 + b_2^2 + \dots + b_m^2).$$

If we write

$$a_1 = f^{(1)}(\xi_1)\{m(\delta_1)\}^{\frac{1}{2}},\ a_2 = f^{(1)}(\xi_2)\{m(\delta_2)\}^{\frac{1}{2}}, \dots a_m = f^{(1)}(\xi_m)\{m(\delta_m)\}^{\frac{1}{2}};$$

$$b_1 = f^{(2)}(\xi_1)\{m(\delta_1)\}^{\frac{1}{2}},\ b_2 = f^{(2)}(\xi_2)\{m(\delta_2)\}^{\frac{1}{2}}, \dots b_m = f^{(2)}(\xi_m)\{m(\delta_m)\}^{\frac{1}{2}},$$

we have

$$\left\{\sum_{r=1}^{r=m} f^{(1)}(\xi_r) f^{(2)}(\xi_r)\, m(\delta_r)\right\}^2 \leq \sum_{r=1}^{r=m} [\{f^{(1)}(\xi_r)\}^2 m(\delta_r)] \sum_{r=1}^{r=m} [\{f^{(2)}(\xi_r)\}^2 m(\delta_r)].$$

Employing the theorem, we have at once the Schwarzian inequality (see § 396),

$$\left\{\int_a^b f^{(1)}(x) f^{(2)}(x)\, dx\right\}^2 \leq \int_a^b \{f^{(1)}(x)\}^2 dx \int_a^b \{f^{(2)}(x)\}^2 dx;$$

since the expressions on the two sides are given to any required degree of approximation by the corresponding expressions in the preceding inequality, provided $\xi_1, \xi_2, \dots \xi_r$ are so chosen that the theorem of § 434 is applicable to the three functions $\{f^{(1)}(x)\}^2$, $\{f^{(2)}(x)\}^2$, $f^{(1)}(x) f^{(2)}(x)$.

The following more general theorems* may be obtained in a similar manner:

$$\left|\int_a^b f^{(1)}(x) f^{(2)}(x)\, dx\right| \leqq \left[\int_a^b \{f^{(1)}(x)\}^p\, dx\right]^{1/p} \left[\int_a^b \{f^{(2)}(x)\}^q\, dx\right]^{1/q},$$

where p and q are any positive numbers such that $1/p + 1/q = 1$; and

$$\left\{\int_a^b \left| f^{(1)}(x) + f^{(2)}(x)\right|^p dx\right\}^{1/p} \leqq \left[\int_a^b \left|f^{(1)}(x)\right|^p dx\right]^{1/p} + \left[\int_a^b \left|f^{(2)}(x)\right|^p dx\right]^{1/p},$$

where $p > 1$.

These are derived from the known inequalities

$$\left|\sum_{r=1}^{r=m} a_r b_r\right| \leqq \left[\sum_{r=1}^{r=m} |a_r|^p\right]^{1/p} \left[\sum_{r=1}^{r=m} |b_r|^q\right]^{1/q},$$

$$\left\{\sum_{r=1}^{r=m} |a_r + b_r|^p\right\}^{1/p} \leqq \left[\sum_{r=1}^{r=m} |a_r|^p\right]^{1/p} + \left[\sum_{r=1}^{r=m} |b_r|^p\right]^{1/p},$$

where p and q are positive numbers such that $1/p + 1/q = 1$.

In the case of the first inequality we take $a_r = f^{(1)}(\xi_r)\{m(\delta_r)\}^{1/p}$, and $b_r = f^{(2)}(\xi_r)\{m(\delta_r)\}^{1/q}$.

This method has been applied† by Lebesgue to obtain a proof of the second mean value theorem (§ 423).

435¹. The above theorems hold good when the integrations are taken over any bounded measurable set E instead of over the interval, or cell, (a, b). For we may suppose the functions $f^{(1)}(x)$, $f^{(2)}(x)$ to have the value zero at all points of the complement of E relatively to an interval, or cell, which contains E. In case E is unbounded, but of finite measure, it is the outer limiting set of a sequence of bounded measurable sets $\{E_n\}$, each of which contains the next. The theorems may be applied to integrals taken over E_n, and then the limits of the integrals, as $n \sim \infty$, may be taken; from § 392 it then follows that the theorems hold good when the integration is taken over E. In case E is measurable, but of infinite measure, employing the definition of an integral over such a set E, given in § 437, we can take E to be the outer limiting set of a sequence $\{E_n\}$ of bounded measurable sets, of which E is the outer limiting set. It then follows, applying the theorems above to integration over the sets E_n, and taking the limit, as $n \sim \infty$, that, when the integrals on the right-hand side of the inequalities exist, the theorems still hold good. It has thus been proved that, if E be a measurable set, bounded or unbounded, such that the integrals on the right-hand side exist,

---

* See F. Riesz, *Math. Annalen*, vol. LXIX (1910), p. 456.

† *Annales de Toulouse* (3), vol. I (1909), p. 36.

$$\left\{\int_{(E)} f^{(1)}(x) f^{(2)}(x)\,dx\right\}^2 \leqq \int_{(E)} \{f^{(1)}(x)\}^2\,dx \int_{(E)} \{f^{(2)}(x)\}^2\,dx \quad \ldots(1).$$

$$\left|\int_{(E)} f^{(1)}(x) f^{(2)}(x)\,dx\right| \leqq \left[\int_{(E)} |f^{(1)}(x)|^p\,dx\right]^{1/p} \left[\int_{(E)} |f^{(2)}(x)|^q\,dx\right]^{1/q} \quad \ldots\ldots\ldots(2).$$

where p and q are any positive numbers such that $1/p + 1/q = 1$;

$$\left\{\int_{(E)} |f^{(1)}(x) + f^{(2)}(x)|^p\,dx\right\}^{1/p} \leqq \left[\int_{(E)} |f^{(1)}(x)|^p\,dx\right]^{1/p} + \left[\int_{(E)} |f^{(2)}(x)|^p\,dx\right]^{1/p} \quad \ldots\ldots\ldots(3),$$

where $p > 1$.

The following relations[*], which are of importance in view of applications, hold also whether $m(E)$ be finite or not:

$$\int_{(E)} |f^{(1)}(x) + f^{(2)}(x)|^k\,dx \leqq \lambda_k \left\{\int_{(E)} |f^{(1)}(x)|^k\,dx + \int_{(E)} |f^{(2)}(x)|^k\,dx\right\} \quad \ldots\ldots\ldots(4),$$

where $k > 0$, and λ_k has the value 2^{k-1} or 1, according as $k > 1$, or $0 < k \leqq 1$;

$$\left|\int_{(E)} |f^{(1)}(x)|^k\,dx - \int_{(E)} |f^{(2)}(x)|^k\,dx\right|$$
$$\leqq k \left\{\int_{(E)} |f^{(1)}(x) - f^{(2)}(x)|^k\,dx\right\}^{1/k}$$
$$\times \left[\left\{\int_{(E)} |f^{(1)}(x)|^k\,dx\right\}^{(k-1)/k} + \left\{\int_{(E)} |f^{(2)}(x)|^k\,dx\right\}^{(k-1)/k}\right] \quad (5),$$

where $k > 1$.

The inequality (4) follows from the known theorems that,

$$(a + b)^k \leqq 2^{k-1}(a^k + b^k),$$

when $k > 1$, and that $(a + b)^k \leqq a^k + b^k$, when $0 < k \leqq 1$; where a and b are positive numbers, by letting $a = |f^{(1)}(x)|$, $b = |f^{(2)}(x)|$, and observing that $|f^{(1)}(x) + f^{(2)}(x)| \leqq |f^{(1)}(x)| + |f^{(2)}(x)|$.

In order to prove (5), we observe that, if a and b are positive numbers such that $a > b$, we have

$$\frac{a^k - b^k}{a - b} = k\,[b + \theta(a - b)]^{k-1},$$

where θ is a number within the interval $(0, 1)$, (see § 262). Hence

$$(a^k - b^k)/(a - b)$$

is less than ka^{k-1}, or kb^{k-1}, according as $k \gtrless 1$; and it is in either case less than $k(a^{k-1} + b^{k-1})$, provided a and b are unequal, whichever be the greater of the two, for $k > 0$. We have accordingly, writing $a = |A|$, $b = |B|$, and observing that $|A - B| \geqq |a - b|$, the inequality

---

[*] See Hobson, *Journal of Lond. Math. Soc.* vol. I (1926), p. 213.

$$\big|\,|A|^k - |B|^k\,\big| \leq k\,|A - B|\,\{|A|^{k-1} + |B|^{k-1}\}.$$

From this inequality it follows that

$$\int_{(E)} \big|\,|f^{(1)}(x)|^k - |f^{(2)}(x)|^k\,\big|\,dx$$

$$\leq k \int_{(E)} |f^{(1)}(x) - f^{(2)}(x)|\,[|f^{(1)}(x)|^{k-1} + |f^{(2)}(x)|^{k-1}]\,dx.$$

Also, if $k > 1$, employing (2), the expression on the right-hand side does not exceed

$$k\left[\left\{\int_{(E)} |f^{(1)}(x) - f^{(2)}(x)|^k\right\}^{1/k}\right]\left[\left\{\int_{(E)} |f^{(1)}(x)|^{k'(k-1)}\right\}^{1/k'}\right.$$
$$\left. + \left\{\int_{(E)} |f^{(2)}(x)|^{k'(k-1)}\right\}^{1/k'}\right],$$

where $1/k + 1/k' = 1$. From this, the inequality (5) follows immediately.

436. Another method has been indicated* by F. Riesz of expressing the L-integrals of functions approximately by finite sums, in such a manner that the above extensions of known inequalities to the case of integrals can be carried out.

It will be sufficient to consider the case of the summable functions $f^{(1)}(x)$, $f^{(2)}(x)$, as the extension to the case of any finite number of such functions will then be obvious

It has been seen, in §§ 385–388, that the integrals $\int_a^b f^{(1)}(x)\,dx$, $\int_a^b f^{(2)}(x)\,dx$ differ respectively from two finite sums

$$a_1 m(e_1) + a_2 m(e_2) + \ldots + a_p m(e_p).$$
$$a_1 m(e_1') + a_2 m(e_2') + \ldots + a_p m(e_p').$$

by less than an arbitrarily chosen positive number, provided the increasing set of numbers $a_1, a_2, \ldots a_p$ is properly chosen, and e_r, e_r' are the sets of points for which $a_r \leq f^{(1)}(x) < a_{r+1}$; $a_r \leqq f^{(2)}(x) < a_{r-1}$, respectively.

If we denote by E_{rs} the set $D(e_r, e_s')$, we have

$$e_r = \sum_{s=1}^{s=p} E_{rs}, \quad e_s' = \sum_{r=1}^{r=p} E_{rs};$$

all the sets E_{rs} can be arranged in order as a single sequence $\{F_q\}$, and the two finite sums then take the forms

$$A_1 m(F_1) + A_2 m(F_2) + \ldots + A_\lambda m(F_\lambda),$$
$$B_1 m(F_1) + B_2 m(F_2) + \ldots + B_\lambda m(F_\lambda),$$

where F_q is the set E_{rs}, and $A_q = a_r$, $B_q = a_s$. In the set F_q, we have

$$a_r \leqq f^{(1)}(x) < a_{r+1}, \quad a_s \leqq f^{(2)}(x) < a_{s+1}.$$

---

* *Loc. cit.*

It is clear that $A_1 B_1 m(F_1) + A_2 B_2 m(F_2) + \dots + A_k B_k m(F_k)$ is an approximation to $\int_a^b f^{(1)}(x) f^{(2)}(x)\,dx$.

LEBESGUE INTEGRALS OVER AN INFINITE FIELD

437. The definition, in §§ 385–388, of the L-integral of a summable function $f(x)$ over a measurable set E, is applicable when E is unbounded, provided it have a finite measure, in accordance with the definition, given in § 134, of the measure of such a set. We proceed to consider the case in which E is a set such that the portion of it in each finite cell, or interval, is measurable, but in which E has not a finite measure. The set E is then said to be measurable, but of infinite measure. Let $f(x)$ be a function, defined over the set E, and everywhere ≥ 0; and let it be assumed that $f(x)$ is summable in every finite cell, or interval, Δ.

If $\int_{(\Delta)} f(x)\,dx$ converges to a definite number, as the distances from the origin of all the boundaries of Δ increase indefinitely, in any manner, that number is said to define the L-integral $\int_{(E)} f(x)\,dx$, of $f(x)$, over the unbounded set E.

If the set E be in p dimensions, $\int_{(\Delta)} f(x)\,dx$ is the measure of the $(p+1)$-dimensional set of points $(x^{(1)}, x^{(2)}, \dots x^{(p)}, y)$, where

$$a^{(r)} \leq x^{(r)} \leq b^{(r)},$$

for $r = 1, 2, 3, \dots p$, and $0 < y \leq f(x^{(1)}, x^{(2)}, \dots x^{(p)})$; the cell Δ being $(a^{(1)}, a^{(2)}, \dots a^{(p)}; b^{(1)}, b^{(2)}, \dots b^{(p)})$. At points x, not in E, we take $f(x) = 0$

It is clear that, when a sequence of cells $\{\Delta_n\}$ is taken, each one of which is contained in the next, and such that, as we progress in the sequence, all the numbers $b^{(1)}, b^{(2)}, \dots b^{(p)}$ become indefinitely great, and are positive, whereas the numbers $a^{(1)}, a^{(2)}, \dots a^{(p)}$ become indefinitely great, and are negative, the values of $\int_{(\Delta_n)} f(x)$ form a monotone sequence. Assuming that this sequence has a finite upper limit, that limit, $\int_{(E)} f(x)\,dx$, is the measure of the $(p+1)$-dimensional unbounded set for which

$$-\infty < x^{(r)} < \infty, \quad (r = 1, 2, 3, \dots p), \quad 0 < y \leq f(x);$$

it being assumed that $f(x) = 0$ at all points that are not in E.

It can be shewn that the limit $\int_{(E)} f(x)\,dx$, when it exists, has its value independent of the particular sequence $\{\Delta_n\}$ of cells, each of which is contained in the next. For we may consider two such sequences $\{\Delta_n^{(1)}\}$, $\{\Delta_n^{(2)}\}$

of cells. For a value of n there is a value of n' such that $\Delta_n^{(1)}$ is contained in $\Delta_{n'}^{(2)}$, and thus

$$\int_{(\Delta_n^{(1)})} f(x)\,dx \leq \int_{(\Delta_{n'}^{(2)})} f(x)\,dx.$$

Since n' must increase indefinitely with n, we have

$$\lim_{n\sim\infty} \int_{(\Delta_n^{(1)})} f(x)\,dx \leq \lim_{n\sim\infty} \int_{(\Delta_n^{(2)})} f(x)\,dx.$$

Since $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$ may clearly be interchanged in this relation, we see that

$$\lim_{n\sim\infty} \int_{(\Delta_n^{(2)})} f(x)\,dx = \lim_{n\sim\infty} \int_{(\Delta_n^{(1)})} f(x)\,dx.$$

If E_n denotes the set of points of E which is contained in Δ_n, $\{E_n\}$ is a sequence of bounded measurable sets, each of which is contained in the next, and of which E is the outer limiting set. Thus $\int_{(E)} f(x)\,dx$ is the limit of the sequence of integrals of $f(x)$ taken over the bounded and measurable sets $\{E_n\}$, of which E is the outer limiting set.

If $f(x)$ is not ≥ 0, at all points of E, it may be expressed as the difference $f^+(x) - f^-(x)$ of two functions $f^+(x)$, $f^-(x)$, each of which is ≥ 0: where $f^-(x) = 0$, at any point x at which $f(x)$ is positive.

In case the two integrals $\int_{(E)} f^+(x)\,dx$, $\int_{(E)} f^-(x)\,dx$ both exist, in accordance with the above definition, the L-integral $\int_{(E)} f(x)\,dx$ is defined to be the value of $\int_{(E)} f^+(x)\,dx - \int_{(E)} f^-(x)\,dx$, and $f(x)$ is said to have an absolutely convergent L-integral over the unbounded set E. The existence of the absolutely convergent integral entails the existence of the integral $\int_{(E)} |f(x)|\,dx$. We have then, as before,

$$\int_{(E)} f(x)\,dx = \lim_{n\sim\infty} \int_{(E_n)} f(x)\,dx.$$

It may however happen that, although $\int_{(\Delta_n)} f^+(x)\,dx$, $\int_{(\Delta_n)} f^-(x)\,dx$ become indefinitely great as the cell, or interval, Δ_n becomes indefinitely great, $\int_{(\Delta_n)} \{f^+(x) - f^-(x)\}\,dx$ converges to a finite limit, independent of the particular sequence $\{\Delta_n\}$. In that case, the limit is said to define the non-absolutely convergent L-integral of $f(x)$ over the unbounded set E. When $\int_{(E)} f(x)\,dx$ exists as a non-absolutely convergent integral,

$$\int_{(E)} |f(x)|\,dx$$

does not exist, because $\int_{(E)} \{f^+(x) + f^-(x)\}\,dx$ does not converge.

438. In the case of integration over a linear interval $\int_a^\infty f(x)\,dx$ is the limit, as $b \sim \infty$, of $\int_a^b f(x)\,dx$; it being assumed that $f(x)$ is summable in (a, b), for all values of b ($> a$). If $\int_a^\infty |f(x)|\,dx$ exists, the integral

$$\int_a^\infty f(x)\,dx$$

is said to be an absolutely convergent L-integral over the unbounded interval (a, ∞) If $\int_a^\infty |f(x)|\,dx$ is not finite, $\int_a^\infty f(x)\,dx$ is said to be a non-absolutely convergent L-integral over the unbounded interval (a, ∞).

If both $\int_0^\infty f(x)\,dx$ and $\int_{-\infty}^0 f(x)\,dx$ exist, their sum defines the value of $\int_{-\infty}^\infty f(x)\,dx$, which is absolutely or non-absolutely convergent over $(-\infty, \infty)$, according as $\int_{-\infty}^\infty |f(x)|\,dx$ is finite or not.

The following theorem provides a criterion for the existence of an L-integral over an unbounded interval which is frequently of use:

If $f(x)$ and $f(x)\,\phi(x)$ be both summable in every interval (a, x), and if $\int_a^\infty f(x)\,dx$ exists, and $\phi(x)$ be monotone and bounded in (X, ∞), for some value of X ($\geqq a$), then $\int_a^\infty f(x)\,\phi(x)\,dx$ exists.

A value ξ, of x, can be so determined that $\left|\int_\xi^{\xi+h} f(x)\,dx\right| < \epsilon$, for all positive values of h, where ϵ is an assigned positive number; and this value of ξ can be chosen to be $\geq X$. We have then, by the second mean value theorem,

$$\int_\xi^{\xi+h} f(x)\,\phi(x)\,dx = \phi(\xi)\int_\xi^{\xi+\theta h} f(x)\,dx + \phi(\xi+h)\int_{\xi+\theta h}^{\xi+h} f(x)\,dx,$$

where θ has some value in the interval $(0, 1)$. If $|\phi(x)| < \kappa$, for every value of x in the interval (X, ∞), we have

$$\left|\int_\xi^{\xi+h} f(x)\,\phi(x)\,dx\right| < 2\kappa\epsilon;$$

and since ϵ is arbitrary, the integral $\int_a^\infty f(x)\,\phi(x)\,dx$ exists.

A similar criterion for the existence of an absolutely convergent integral is the following:

If[*] *$f(x)$ and $f(x)\,\phi(x)$ be both summable in every interval (a, x), and if*

---

[*] See Riemann s *Werke*, 1st ed , p. 229; also Pringsheim, *Math. Annalen*, vol. XXXVII (1890), p. 591.

$\int_a^{\infty} f(x)\,dx$ *exists as an absolutely convergent L-integral, and* $|\phi(x)|$ *be bounded in* (X, ∞), *for some value of* X $(\geqq a)$, *then the integral*

$$\int_a^{\infty} f(x)\,\phi(x)\,dx$$

exists, as an absolutely convergent L-integral.

For, since $\int_a^{\infty} f(x)\,dx$ is absolutely convergent, a number ξ $(\geqq X)$ can be so determined that $\int_{\xi}^{\xi+h} |f(x)|\,dx < \epsilon$, for all positive values of h; where ϵ is any assigned positive number.

We have then

$$\left|\int_{\xi}^{\xi+h} f(x)\,\phi(x)\,dx\right| \leqq \int_{\xi}^{\xi+h} |f(x)\,\phi(x)|\,dx < \kappa \int_{\xi}^{\xi+h} |f(x)|\,dx < \kappa\epsilon,$$

where κ is the upper boundary of $|\phi(x)|$ in (ξ, ∞). Since ϵ is arbitrary, the condition for the existence of $\int_a^{\infty} |f(x)\,\phi(x)|\,dx$ is satisfied.

439. An important class of integrals over an unbounded interval, of which the convergence is not necessarily absolute, is that of the integrals

$$\int_0^{\infty} \phi(x)\sin x\,dx, \quad \int_0^{\infty} \phi(x)\cos x\,dx,$$

where $\phi(x)$ is monotone, from and after some fixed value of x, and converges to zero, as x is indefinitely increased.

We have $\int_{x_1}^{x_2} \phi(x)\sin x\,dx = \phi(x_1)\int_{x_1}^{\xi} \sin x\,dx + \phi(x_2)\int_{\xi}^{x_2} \sin x\,dx$, where x_1 is so large that $\phi(x)$ is monotone for $x \geqq x_1$, and ξ is some number in the interval (x_1, x_2).

From this we have

$$\left|\int_{x_1}^{x_2} \phi(x)\sin x\,dx\right| \leqq 2\,|\phi(x_1)| + 2\,|\phi(x_2)|.$$

Since $|\phi(x)|$ converges to zero, x_1 may be so chosen that $|\phi(x_1)|$ and $|\phi(x_2)|$ are both arbitrarily small; hence the condition is satisfied for the existence of $\int_0^{\infty} \phi(x)\sin x\,dx$, as an L-integral which is not necessarily absolutely convergent. The case of the second integral may be treated in the same manner.

CHANGE OF THE INDEPENDENT VARIABLE IN A LEBESGUE INTEGRAL

440. The transformation to a new variable, of the integral $\int_a^b f(x)\,dx$, of a function $f(x)$ that is summable in the linear interval (a, b) will now be considered. The following theorem* will be established:

* See Lebesgue, *Annales de Toulouse* (3), vol. I (1909), p. 44.

If $f(x)$ be summable in the linear interval (a, b), and if $\phi(\xi)$ be a monotone function of ξ which is an indefinite integral $a + \int_{\alpha}^{\xi} \chi(\xi)\, d\xi$, where the interval (α, β), of ξ, is made to correspond to the interval (a, b), of x, by means of the relation $x = \phi(\xi)$, then $\int_a^x f(x)\, dx = \int_\alpha^\xi F(\xi)\, \chi(\xi)\, d\xi$, where $F(\xi)$ denotes the function $f\{\phi(\xi)\}$.

We may assume, without loss of generality, that $\phi(\xi)$ is non-diminishing in (α, β). It has been shewn, in § 407, that the necessary and sufficient condition that the monotone function $\phi(\xi)$ should be an indefinite integral is that, to any set of points of measure zero on the ξ-segment, there should correspond a set of points of measure zero on the x-segment.

The above theorem is equivalent* to the statement that, when the condition here referred to is satisfied, then the equality

$$\int_a^x f(x)\, dx = \int_\alpha^\xi F(\xi)\, \chi(\xi)\, d\xi$$

holds for every summable function $f(x)$.

Let $\Delta_n^{(x)}$ denote a set of non-overlapping intervals in (a, b), and $\Delta_n^{(\xi)}$ the corresponding set of intervals in (α, β); we have then

$$m(\Delta_n^{(x)}) = \int_{(\Delta_n^{(\xi)})} \chi(\xi)\, d\xi$$

If $E^{(x)}$ be the inner limiting set of a sequence of sets of intervals, each of which sets contains the next, there corresponds to $E^{(x)}$ a set $E^{(\xi)}$, the inner limiting set of the corresponding sequence of intervals in the ξ-segment. We have then

$$m(E^{(x)}) = \lim_{n \sim \infty} m(\Delta_n^{(x)}) = \lim_{n \sim \infty} \int_{(\Delta_n^{(\xi)})} \chi(\xi)\, d\xi = \int_{(E^{(\xi)})} \chi(\xi)\, d\xi.$$

If $m(E^{(x)}) = 0$, it follows that $\chi(\xi) = 0$, at all points of $E^{(\xi)}$, with the possible exception of those which belong to a component of measure zero.

The set of points on the ξ-segment that corresponds to a set of points of measure zero on the x-segment has not necessarily the measure zero, and is conceivably not a measurable set; but in any case that component of the set at which $\chi(\xi)$ is not zero is of measure zero.

A set of points on the x-segment which is measurable (B) corresponds to a set on the ξ-segment that is also measurable (B). Any measurable set $E^{(x)}$, on the x-segment, contains a set $E_1^{(x)}$, which is measurable (B), and such that $E^{(x)} - E_1^{(x)}$ has measure zero. If $E^{(x)}$, $E_1^{(x)}$ correspond to $E^{(\xi)}$, $E_1^{(\xi)}$, on the ξ-segment, $E^{(\xi)}$ contains the measurable set $E_1^{(\xi)}$, and at all points of the set $E^{(\xi)} - E_1^{(\xi)}$, with the exception of a component of measure zero, the condition $\chi(\xi) = 0$ is satisfied. Thus $E^{(\xi)}$ is the sum of a measurable set, and of a set of points in which $\chi(\xi) = 0$.

---

* See Hahn, *Monatshefte fur Math. u. Physik*, vol. XXIII (1912), p. 167.

If $E^{(x)}$ is the set of points at which $f(x) > A$, $E^{(\xi)}$ is the set of points at which $F(\xi) > A$. If $F_1(\xi)$ is a function defined only at those points of the interval (α, β) in which $\chi(\xi) \neq 0$, and is equal to $F(\xi)$ at the points for which it is defined, the set of points at which $F_1(\xi) > A$ is measurable. It follows that $F_1(\xi)\,\chi(\xi)$ is a measurable function, defined over that set of points at which $\chi(\xi) \neq 0$; and thus $F(\xi)\,\chi(\xi)$ is measurable over (α, β)

Let us first consider the case in which $f(x)$ is continuous in (a, b): then $F(\xi)$ is continuous in (α, β) Let a net $(a, x_1, x_2, \ldots x_{n-1}, b)$ be fitted on to (a, b); and let $(\alpha, \xi_1, \xi_2, \ldots \xi_{n-1}, \beta)$ be corresponding points of (α, β). If U_r, L_r are the upper and lower boundaries of $f(x)$ in (x_{r-1}, x_r), or of $F(\xi)$ in (ξ_{r-1}, ξ_r), we have

$$U_r \int_{\xi_{r-1}}^{\xi_r} \chi(\xi)\, d\xi \geqq \int_{\xi_{r-1}}^{\xi_r} F(\xi)\,\chi(\xi)\, d\xi \geqq L_r \int_{\xi_{r-1}}^{\xi_r} \chi(\xi)\, d\xi.$$

Therefore

$$\sum_{r=1}^{r=n} U_r (x_r - x_{r-1}) \geqq \int_{\alpha}^{\beta} F(\xi)\,\chi(\xi)\, d\xi \geqq \sum_{r=1}^{r=n} L_r (x_r - x_{r-1}).$$

If the net be one of a system of nets fitted on to (a, b), the two sums converge to $\int_a^b f(x)\, dx$, when the nets of the system are taken successively. It then follows that

$$\int_{\alpha}^{\beta} F(\xi)\,\chi(\xi)\, d\xi = \int_a^b f(x)\, dx.$$

Since b and β may be replaced by x and ξ, the theorem has been established for the case of a continuous function $f(x)$

Next, let $f(x)$ be any summable function, bounded in (a, b). A continuous function $f_\epsilon(x)$ can be so determined that $\int_a^b |f(x) - f_\epsilon(x)|\, dx < \epsilon^2$, and so that the upper boundary of $|f_\epsilon(x)|$ does not exceed that of $|f(x)|$ (see § 430). Denoting by $F_\epsilon(\xi)$ the function that is equivalent to $f_\epsilon(x)$, we have

$$\left| \int_{\alpha}^{\beta} \{F(\xi) - F_\epsilon(\xi)\}\,\chi(\xi)\, d\xi \right| < \epsilon \int_{\alpha}^{\beta} \chi(\xi)\, d\xi + 2\overline{U} \int_{(H_1)} \chi(\xi)\, d\xi:$$

where $\overline{U}$ is the upper boundary of $|f(x)|$ in (a, b), and H_1 is that set of points at which $\chi(\xi) \neq 0$, and which is a component of the set that corresponds to the set H on the x-segment, in which $|f(x) - f_\epsilon(x)| > \epsilon$. The measure of H cannot exceed ϵ; and if we assign to ϵ successively the values in a diminishing sequence that converges to zero, the set $H^{(\epsilon_r)}$ which corresponds to a value ϵ_r in the sequence contains the set $H^{(\epsilon_{r-1})}$. The inner limiting set of the sequence $\{H^{(\epsilon_r)}\}$ has the measure zero, and therefore the inner limiting set of the corresponding sequence $\{H_1^{(\epsilon_r)}\}$ has the measure zero, as it contains no points at which $\chi(\xi) = 0$.

It follows that $m(H_1) \sim 0$, as $\epsilon \sim 0$; and thus we have

$$\int_{\alpha}^{\beta} F(\xi)\, \chi(\xi)\, d\xi = \lim_{\epsilon \sim 0} \int_{\alpha}^{\beta} F_{\epsilon}(\xi)\, \chi(\xi)\, d\xi = \lim_{\epsilon \sim 0} \int_a^b f_{\epsilon}(x)\, dx.$$

It has now been shewn that

$$\int_{\alpha}^{\beta} F(\xi)\, \chi(\xi)\, d\xi = \int_a^b f(x)\, dx,$$

where $f(x)$ is any bounded summable function.

Lastly, let $f(x)$ be unbounded and summable; since it can be expressed as the difference of two non-negative summable functions, it will be sufficient to consider the case in which $f(x) \geq 0$, in (a, b). Let $f_N(x) = f(x)$, at points where $f(x) \leq N$, and let $f_N(x) = N$, at points where $f(x) > N$; the positive number N being arbitrary. We have then

$$\int_a^b f_N(x)\, dx = \int_{\alpha}^{\beta} F_N(\xi)\, \chi(\xi)\, d\xi,$$

where $F_N(\xi)$ is equivalent to $f_N(x)$. Since the integral on the left-hand side is non-diminishing as N is increased, and since it has the finite limit $\int_a^b f(x)\, dx$, as $N \sim \infty$, it follows from the theorem of § 399 that

$$\int_{\alpha}^{\beta} F(\xi)\, \chi(\xi)\, d\xi = \int_a^b f(x)\, dx.$$

We may replace b by any point x, in the interval (a, b), and β by the point in (α, β) that corresponds to x.

If E be any measurable set in (a, b), and E_1 be the measurable component of that set in (α, β) which corresponds to E, and in which $\chi(\xi) \neq 0$, we may replace $f(x)$ by a function which has the same value as $f(x)$ in E, and is zero in $C(E)$; we have then

$$\int_{(E)} f(x)\, dx = \int_{(E_1)} F(\xi)\, \chi(\xi)\, d\xi.$$

In accordance with the theorem of § 405, since $x = a + \int_{\alpha}^{\xi} \chi(\xi)\, d\xi$, we have $\dfrac{dx}{d\xi} = \chi(\xi)$, for almost all values of ξ in (α, β). Hence the above theorem may be written in the form

$$\int_a^b f(x)\, dx = \int_{\alpha}^{\beta} F(\xi)\, \frac{dx}{d\xi}\, d\xi;$$

it being assumed that $\dfrac{dx}{d\xi}$ is integrable (L), in (α, β).

The more general case of substitution in which the function $\phi(\xi)$ is not necessarily monotone but is absolutely continuous, and is consequently an indefinite integral, has been treated* by de la Vallée Poussin. He has shewn that the equality

$$\int_a^b f(x)\,dx = \int_\alpha^\beta F(\xi)\,\phi'(\xi)\,d\xi,$$

where $a = \phi(\alpha)$, $b = \phi(\beta)$, $F(\xi) = f\{\phi(\xi)\}$, is valid provided either (1), $\overline{F}\{\phi(\xi)\}$ is absolutely continuous, where $\overline{F}(x)$ denotes $\int_a^x f(x)\,dx$; or (2), if $F(\xi)\,\phi'(\xi)$ is summable. The case has been considered by Fichtenholz in which $\phi(\xi)$ is not necessarily absolutely continuous, but is continuous and monotone He has shewn† that the formula is still valid when the function $\overline{F}\{\phi(\xi)\}$ is absolutely continuous.

441. The extension of the theorem of § 373 to the case in which the function $f(x, y)$ is any summable function will here be considered.

It will be assumed, as in § 372, that a (1, 1) correspondence between the points of a perfect connex domain H, in the plane of (x, y), and the corresponding domain $\overline{H}$, in the plane of (ξ, η), is defined by $x = f_1(\xi, \eta)$, $y = f_2(\xi, \eta)$, where the partial differential coefficients of f_1 and f_2 are continuous in $\overline{H}$, and the Jacobian $\frac{\partial(f_1, f_2)}{\partial(\xi, \eta)} \equiv J$ does not vanish in $\overline{H}$. It was shewn, in § 372, that, to a closed set h, in H, of content zero, there corresponds a closed set $\overline{h}$, in $\overline{H}$, of content zero.

If δ denote a closed cell in H, and $\overline{\delta}$ the corresponding closed domain in $\overline{H}$, by taking $f(x, y) = 1$, in H, we have, in accordance with the result obtained in § 373,

$$m(\delta) = \int_{(\delta)} f(x, y)\,d(x, y) = \int_{(\overline{\delta})} F(\xi, \eta)\,|J|\,d(\xi, \eta) = km(\overline{\delta}),$$

where k is some number between the upper and lower limits of $|J|$ in $\overline{\delta}$.

If G be any connex perfect domain of which the frontier has its plane content zero, G being contained in H, and if $\overline{G}$, in the plane of (ξ, η), correspond to G, we see that, $|J|$ being continuous and never vanishing, it has a lower limit $1/\lambda$, in $\overline{G}$, where $\lambda > 0$; therefore $m(\overline{\delta}) < \lambda m(\delta)$, for every cell δ, contained in G.

If Δ be a set of non-overlapping cells, contained in G, and $\overline{\Delta}$ be the set of domains in $\overline{G}$ that corresponds to Δ, we have $m(\overline{\Delta}) < \lambda m(\Delta)$.

Let E be any measurable set of points in G, then E is contained in a set E_1, of measure $m(E)$, which is the inner limiting set of a sequence $\{\Delta^{(n)}\}$, of sets of non-overlapping cells. Let $\{\overline{\Delta}^{(n)}\}$ be the corresponding

---

[*] *Cours d'Analyse infinitésimale*, vol I (1914), pp. 280–284, also *Trans. Amer. Math. Soc.* vol. XVI (1915), p. 466. For a critical remark on the proof in this memoir, see Fichtenholz, *Bulletin de l'Acad. roy de Belgique* (1922), p. 441.

[†] *Loc. cit.* p. 443.

sequence of non-overlapping domains in $\overline{G}$. The set $E_1 - E - F$, which has measure zero, is contained in a set F_1, also of measure zero, which is the inner limiting set of a sequence $\{\Delta'^{(n)}\}$ of sets of cells. Let $\overline{F}_1$ be the set, in G, which corresponds to F_1; then

$$m(\overline{F}_1) = \lim_{n\sim\infty} m(\overline{\Delta}'^{(n)}) \leq \lambda \lim_{n\sim\infty} m(\Delta'^{(n)});$$

hence $m(\overline{F}_1) = 0$. The set E is contained in E_1, and contains $E_1 - F_1$, and hence the set $\overline{E}$, corresponding to E, is contained in $\overline{E}_1$, and contains $\overline{E}_1 - \overline{F}_1$. Since $m(\overline{F}_1) = 0$, and $\overline{E}_1$ is measurable it follows that $\overline{E}$ is measurable, and

$$m(\overline{E}) = m(\overline{E}_1) \leq \lambda m(E_1) = \lambda m(E).$$

It has thus been shewn that the set $\overline{E}$, in $\overline{G}$, that corresponds to a measurable set E, in G, is measurable, and that its measure does not exceed $\lambda m(E)$.

If $f(x, y)$ be a measurable function, defined over the set G, the function $F(\xi, \eta)$ which has the same value at a point (ξ, η), of $\overline{G}$, as $f(x, y)$ has at the point (x, y) that corresponds to (ξ, η) is measurable in $\overline{G}$.

First, let $f(x, y)$ be bounded in G; then a continuous function $\phi_\epsilon(x, y)$ can be so determined (see § 430) that

$$\int_{(G)} | f(x, y) - \phi_\epsilon(x, y) | \, d(x, y) \leq \epsilon^2.$$

In a set L, of points of measure $> m(G) - \epsilon$, we have

$$| f(x, y) - \phi_\epsilon(x, y) | \leq \epsilon.$$

Moreover $\phi_\epsilon(x, y)$ can be so defined that its numerical value never exceeds the upper boundary of $| f(x, y) |$ in G.

From the theorem established in § 373, we have

$$\int_{(G)} \phi_\epsilon(x, y) \, d(x, y) = \int_{(\overline{G})} F_\epsilon(\xi, \eta) \, | J | \, d(\xi, \eta),$$

where $F_\epsilon(\xi, \eta) = \phi_\epsilon(x, y)$, at corresponding points.

To estimate the value of

$$\left| \int_{(\overline{G})} F(\xi, \eta) \, | J | \, d(\xi, \eta) - \int_{(\overline{G})} F_\epsilon(\xi, \eta) \, | J | \, d(\xi, \eta) \right|$$

we divide the set $\overline{G}$ into the two parts $\overline{L}$ and $C(\overline{L})$, where, in $\overline{L}$, the condition $| F(\xi, \eta) - F_\epsilon(\xi, \eta) | \leq \epsilon$ is satisfied. The absolute value of the difference is then seen to be

$$< \epsilon \int_{(\overline{G})} | J | \, d(\xi, \eta) + 2\overline{U} \int_{C(\overline{L})} | J | \, d(\xi, \eta);$$

where $\overline{U}$ is the upper boundary of $| f(x, y) |$ in G.

Since $m\{C(\overline{L})\} \leq \lambda m\{C(L)\} < \lambda\epsilon$, we see that this expression is arbitrarily small, if ϵ be taken small enough.

It has thus been shewn that

$$\lim_{\epsilon \sim 0} \int_{(\overline{G})} F_\epsilon(\xi, \eta) \, | J | \, d(\xi, \eta) = \int_{(\overline{G})} F(\xi, \eta) \, | J | \, d(\xi, \eta)$$

and therefore we have

$$\int_{(G)} f(x, y) \, d(x, y) = \int_{(\overline{G})} F(\xi, \eta) \, | J | \, d(\xi, \eta).$$

Next, let $f(x, y)$ be unbounded. Since it can be expressed as the difference of two summable functions, each of which is non-negative, it will be sufficient to assume that $f(x, y) \geq 0$. Let N be an arbitrarily chosen positive number and let $f_N(x, y) = f(x, y)$, at all points where $f(x, y) \leq N$, and elsewhere let $f_N(x, y) = N$. We have then, since $f_N(x, y)$ is bounded,

$$\int_{(G)} f_N(x, y) \, d(x, y) = \int_{(\overline{G})} F_N(\xi, \eta) \, | J | \, d(\xi, \eta),$$

where $F_N(\xi, \eta)$ has the same value as $f_N(x, y)$, at the point (x, y) which corresponds to (ξ, η).

Since $$\lim_{N \sim \infty} \int_{(G)} f_N(x, y) \, d(x, y) = \int_{(G)} f(x, y) \, d(x, y),$$

it follows that $\lim\limits_{N \sim \infty} \int_{(\overline{G})} F_N(\xi, \eta) \, | J | \, d(\xi, \eta)$ exists, and is equal to

$$\int_{(G)} f(x, y) \, d(x, y).$$

Since $F_N(\xi, \eta) \, | J |$ is monotone as N is increased, it follows, by applying the theorem of § 399, that

$$\int_{(\overline{G})} F(\xi, \eta) \, | J | \, d(\xi, \eta) = \int_{(G)} f(x, y) \, d(x, y).$$

The transformation obtained in § 373 has now been extended to the case in which $f(x, y)$ is any function that is summable in G.

The extension, given in § 374, to the case in which, in a closed set of points of content zero, the Jacobian either vanishes or is indefinite, is applicable when $f(x, y)$ is a summable function. Moreover the case in which one of the domains G, $\overline{G}$ is unbounded, or in which both are unbounded, can be considered, as in § 375.

442. A more general treatment of the transformation of a double, or of a multiple, integral has been given by W. H. Young, in which it is not assumed that the Jacobian is necessarily of one sign, or that the correspondence of the points (x, y), (ξ, η) is necessarily a (1, 1) correspondence. He has established the following theorem:

Let $x = x(\xi, \eta)$, $y = y(\xi, \eta)$ be functions of (ξ, η) possessing the property of having all their partial derivatives with respect to (ξ, η) bounded for all values of (ξ, η) in the fundamental rectangle $(\alpha_1, \beta_1; \alpha_2, \beta_2)$; and let A be *the area of the curve in the (x, y)-plane which is the image of the perimeter of this fundamental rectangle. Then*

$$A = \int_{\alpha_1}^{\alpha_2} \left\{ \int_{\beta_1}^{\beta_2} \frac{\partial (x, y)}{\partial (\xi, \eta)} d\eta \right\} d\xi ;$$

where $\frac{\partial x}{\partial \xi}, \frac{\partial x}{\partial \eta}, \frac{\partial y}{\partial \xi}, \frac{\partial y}{\partial \eta}$ *represent any of the partial derivatives of x and y with respect to ξ and η.*

More generally, the same is true, if the partial derivatives are not bounded, provided only the following conditions are satisfied:

(1). *$x(\xi, \eta)$ is an integral with respect to ξ, and an integral with respect to η;*

(2). *$y(\xi, \eta)$ is an integral with respect to ξ, and an integral with respect to η;*

(3). $\frac{\partial y}{\partial \eta}$ *and* $\frac{\partial x}{\partial \eta}$ *are, except for a set of values of η, of measure zero, less than summable functions of η alone, say $\mu(\eta)$ and $M(\eta)$;*

(4). *The same condition as* (3) *is satisfied, with ξ and η interchanged, or more generally, when* $\int \frac{\partial x}{\partial \xi} \mu(\eta) \, d(\xi, \eta)$, $\int \frac{\partial y}{\partial \xi} M(\eta) \, d(\xi, \eta)$ *exist as absolutely convergent integrals.*

For the proof of this theorem reference may be made to a memoir* by W. H. Young.

For the transformation of the integral $f(x, y)$, he has given the following theorem:

Let $x = x(\xi, \eta)$, $y = y(\xi, \eta)$, where the functions are continuous with respect to (ξ, η), in a fundamental cell. Let it be assumed that any cell in the (ξ, η)-space has for its image in the (x, y)-space, a portion of that space whose boundary divides it into two distinct portions, an exterior and an interior. Further let it be assumed that the area of a portion of the (x, y)-space is given by $\int \frac{\partial (x, y)}{\partial (\xi, \eta)} d(\xi, \eta)$, *an absolutely convergent integral, which may be positive, negative, or zero. Then* $\int f(x, y) \, d(x, y)$, *taken over the portion of the (x, y)-space which corresponds to a cell in the (ξ, η)-space, is equal to*

$$\int f\{x(\xi, \eta), y(\xi, \eta)\} \frac{\partial (x, y)}{\partial (\xi, \eta)} d(\xi, \eta)$$

taken over the cell.

For the proof of this theorem reference† may be made to a memoir by W. H. Young.

---

* See *Proc. Lond. Math. Soc.* (2), vol. XVIII (1918), p. 339, "On a formula for an area."

† *Proc. Royal Soc.* vol. XCVI (1920), p. 82.

HARNACK'S DEFINITION OF AN INTEGRAL

443. A definition of the integral of a function $f(x)$, unbounded in the linear interval for which it is defined, was given* by Harnack, which in its original form depended upon the employment of Riemann's integrals in sub-intervals of (a, b) in which such integrals exist. The definition given by Harnack can however, without essential change of form, be extended to the case in which Lebesgue integrals take the place of Riemann integrals.

The set of points of infinite discontinuity of a function $f(x)$, defined for the finite interval (a, b), form a closed set G.

It will be assumed that this closed set G is of content zero. Let G be enclosed within the intervals $\delta_1, \delta_2, \ldots \delta_n$, of a finite set Δ, such that each interval of Δ contains at least one point of G. The complement $C'(\Delta)$ consists of a finite set of intervals, free at their ends, and in their interiors, from points of G.

Let it be assumed that $f(x)$ is integrable (L) in each of the intervals of $C(\Delta)$, and thus that $\int_{C(\Delta)} f(x)\,dx$ exists as an L-integral. The integral of $f(x)$, in (a, b), when it exists, is defined as follows.

Let it be assumed that a number I exists such that, corresponding to each arbitrarily chosen positive number ϵ, a positive number ζ_ϵ can be determined so that

$$\left| I - \int_{C(\Delta)} f(x)\,dx \right| < \epsilon$$

for every finite set Δ of intervals which satisfy the conditions given above, and which is such that $m(\Delta) < \zeta_\epsilon$. The number I is then taken to define the value of

$$\int_a^b f(x)\,dx.$$

It is convenient to define a function $f_\Delta(x)$ which is zero in all points of the intervals Δ, and is equal in value to $f(x)$, at all interior points of $C(\Delta)$.

We have then, in accordance with the above definition,

$$\int_a^b f(x)\,dx = \lim_{m(\Delta)\sim 0} \int_a^b f_\Delta(x)\,dx,$$

the convergence being uniform with respect to all sets Δ.

The necessary and sufficient condition for the existence of the integral $\int_a^b f(x)\,dx$, as a finite number, is that, ϵ being assigned, ζ_ϵ can be so determined that

$$\left| \int_a^b f_\Delta(x)\,dx - \int_a^b f_{\Delta'}(x)\,dx \right| < 2\epsilon$$

---

* *Math Annalen*, vol. XXIV (1884), p. 220. See also Jordan, *Cours d'Analyse*, vol. II (1884), p 50, where a similar definition is given, except that the condition that the set of points of infinite discontinuity should have content zero is omitted.

for every pair of finite sets Δ, Δ' of intervals, such that each interval of either set encloses within it at least one point of G, provided

$$m(\Delta) < \zeta_\epsilon, \quad m(\Delta') < \zeta_\epsilon;$$

and that this condition is satisfied whatever value ϵ may have.

It is easily seen that the above definition is equivalent to that of Cauchy (§ 352), in case the set G consists of a finite number of points

It will be shewn that the above definition is more general than that of Lebesgue, in that it defines integrals which do not necessarily exist in accordance with Lebesgue's definition. It is clearly less general than Lebesgue's definition, in that it applies only to the case in which the set of points of infinite discontinuity of the integrand is of content zero.

As $m(\Delta)$ converges to zero, the number n, of intervals of Δ, will increase indefinitely. If any sequence $\Delta_1, \Delta_2, \Delta_3, \ldots$ corresponding to a sequence $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$, of values of ϵ, which converges to zero, be taken, a sequence

$$\Delta_{n_1}, \Delta_{n_2}, \Delta_{n_3}, \ldots$$

may be chosen out of the given sequence in such a way that Δ_{n_r} contains $\Delta_{n_{r+1}}$, for all values of r For let $C(\Delta_{n_1})$ consist of, say, r_{n_1} intervals (α, β) each contained in some interval $(\bar{\alpha}, \bar{\beta})$ contiguous to G. The number n_2 can be so chosen that it is the smallest integer $(> n_1)$ such that $m(\Delta_{n_2})$ is less than the least of all the $2r_1$ numbers $\alpha - \bar{\alpha}, \bar{\beta} - \beta$. Clearly then (α, β) is within an interval complementary to the set Δ_{n_2} The numbers $n_3, n_4, \ldots$ may then be chosen successively by the same procedure. The sequence $\Delta_{n_1}, \Delta_{n_2}, \ldots$ is therefore such that each set contains the next.

When both $\int_a^b f(x)\,dx$ and $\int_a^b |f(x)|\,dx$ exist, in accordance with the above definition, $\int_a^b f(x)\,dx$ is said to be absolutely convergent. Those integrals which exist in accordance with the above definition, and are not absolutely convergent, will be considered in Chapter VIII.

444. It will be shewn that:

If $\int_a^b f(x)\,dx$ exists, and is absolutely convergent, the integral also exists as an L-integral, and conversely, that, if it exists as an L-integral, and the set of points of infinite discontinuity has content zero, it also exists as an absolutely convergent integral, in accordance with the above definition.

First, let $f(x) \geq 0$, in the interval (a, b). Taking the sequence

$$\Delta_1, \Delta_2, \Delta_3, \ldots$$

so that each set contains the next, we see that the values of

$$\int_a^b f_{\Delta_r}(x)\,dx,$$

for $r = 1, 2, 3, \ldots$, form a monotone non-diminishing sequence of numbers, and thus they either increase indefinitely, or they converge to a limit A. Assume that the latter is the case.

Let $\Delta_1', \Delta_2', \Delta_3', \dots$ be any other sequence of finite sets of intervals which satisfy the conditions in the definition, but not necessarily such that Δ_s' contains Δ'_{s+1}, for all values of s. We have

$$\int_a^b f_{\Delta'_s}(x)\,dx - \int_a^b f_{\Delta_r}(x)\,dx < Pm(\Delta_r):$$

where P is the upper boundary of $f(x)$ in the intervals of $C(\Delta_s')$: for the above difference is less than the integral of $f(x)$ over those intervals that are in Δ_r and also in $C(\Delta_s')$. If ϵ be arbitrarily chosen, and s be fixed, for all sufficiently large values of r we have, on the assumption of the existence of the limit A, of $\int_a^b f_{\Delta_r}(x)\,dx$,

$$\int_a^b f_{\Delta'_s}(x)\,dx - \int_a^b f_{\Delta_r}(x)\,dx < \epsilon,$$

and therefore

$$\int_a^b f_{\Delta'_s}(x)\,dx \leqq A + \epsilon;$$

this holds for each value of s. Similarly, we see that, for a fixed value of r, and for all sufficiently large values of s,

$$\int_a^b f_{\Delta'_s}(x)\,dx > \int_a^b f_{\Delta_r}(x)\,dx - \tfrac{1}{2}\epsilon;$$

and by choosing the fixed value of r so great that

$$\int_a^b f_{\Delta_r}(x)\,dx > A - \tfrac{1}{2}\epsilon,$$

we have

$$\int_a^b f_{\Delta'_s}(x)\,dx > A - \epsilon,$$

for all sufficiently large values of s. As

$$\int_a^b f_{\Delta'_s}(x)\,dx$$

lies in the interval $(A - \epsilon, A + \epsilon)$ for all sufficiently large values of s, and since ϵ is arbitrary, it follows that

$$\int_a^b f_{\Delta'_s}(x)\,dx$$

converges to A. Moreover, since

$$\int_a^b f_{\Delta_r}(x)\,dx - \int_a^b f_{\Delta'_s}(x)\,dx < P'm(\Delta'_s),$$

where P' is the upper boundary of $f(x)$ in $C(\Delta_r)$; and also since

$$\int_a^b f_{\Delta_r}(x)\,dx - \int_a^b f_{\Delta''_{s'}}(x)\,dx < P'm(\Delta''_{s'}),$$

where $\{\Delta''_{s'}\}$ is any other sequence of intervals enclosing G; from these inequalities, we see that, if $m(\Delta'_s)$, $m(\Delta''_{s'})$ are both less than $\frac{1}{2}\epsilon/P'$, the integrals

$$\int_a^b f_{\Delta'_s}(x)\,dx, \quad \int_a^b f_{\Delta''_{s'}}(x)\,dx$$

are both $> A - \epsilon$, and they have been shown to be $< A + \epsilon$; it follows that the integrals differ from one another by less than 2ϵ. Hence the condition for the existence of the integral $\int_a^b f(x)\,dx$ is satisfied, on the assumption that the integrals of $f_{\Delta_r}(x)$ converge to a finite number as $r \sim \infty$.

Thus it has been shown that:

If $f(x)$ be never negative, it is necessary and sufficient for the existence of the integral $\int_a^b f(x)\,dx$ in accordance with Harnack's definition that, if $\Delta_1, \Delta_2, \ldots$ be a sequence of finite sets of intervals each containing the next, such that each interval of each set contains within it a point of G, and such that $m(\Delta_r) \sim 0$, then $\int_a^b f_{\Delta_r}(x)\,dx$ should be less than a fixed finite number, for all the sets Δ_r.

Let it now be assumed that $\int_a^b f(x)\,dx$ exists as an L-integral; we have then

$$\int_a^b f_\Delta(x)\,dx = \int_a^b f(x)\,dx - \int_{(\Delta)} f(x)\,dx;$$

and since (§ 392), $\int_{(\Delta)} f(x)\,dx$ converges to zero, uniformly as $m(\Delta) \sim 0$, we see that $\int_a^b f_\Delta(x)\,dx$ converges to $\int_a^b f(x)\,dx$; and thus the integral also exists in accordance with Harnack's definition. It appears moreover that in this case of a non-negative function the condition that each interval of Δ should contain a point of G is unnecessary.

It has thus been proved that, if $f(x)$ be non-negative, and have an L-integral, it has also an integral in accordance with Harnack's definition, and the two have the same value.

Next, let $f(x)$ have both positive and negative values, in (a, b), and assume that it is summable in the interval.

If $f(x) = f^+(x) - f^-(x)$, $f^+(x)$ and $f^-(x)$ are summable, and therefore they have not only L-integrals but also Harnack integrals. It is clear that

$$\int_a^b f_\Delta(x)\,dx = \int_a^b f_\Delta{}^+(x)\,dx - \int_a^b f_\Delta{}^-(x)\,dx.$$

Now, as $m(\Delta)$ converges to zero, the two integrals on the right-hand side converge to

$$\int_a^b f^+(x)\,dx, \quad \int_a^b f^-(x)\,dx,$$

as has been shewn above, although Δ is such that an interval of it does not necessarily contain a point both of G_1 and of G_2, the sets of points of infinite discontinuity of $f^+(x)$ and of $f^-(x)$.

It follows that $\int_a^b f(x)\,dx$ exists in accordance with Harnack's definition, and is equal to the L-integral.

It follows immediately that $\int_a^b |f(x)|\,dx$ exists as a Harnack integral, and is equal to the sum of the integrals of $f^+(x)$ and $f^-(x)$.

Conversely, let us assume that

$$\int_a^b f(x)\,dx \text{ and } \int_a^b |f(x)|\,dx$$

exist as Harnack integrals.

The points of infinite discontinuity of the two functions $f(x)$, $|f(x)|$ are the same, hence the Harnack integrals of the two functions are the limits of L-integrals taken over sets of intervals that are the same for the two functions. It follows that $|f(x)|+f(x)$, $|f(x)|-f(x)$ have Harnack integrals, the sum and difference of those of $|f(x)|$ and $f(x)$; or the two functions $f^+(x)$, $f^-(x)$ have Harnack integrals.

It is then sufficient to shew that the existence of the Harnack integral of a non-negative function involves the existence of the L-integral of the same function.

Let it therefore be assumed that $f(x) \geqq 0$, and that the Harnack integral of $f(x)$ over (a, b) exists The points of infinite discontinuity of $f(x)$ can be enclosed in a finite set Δ, of intervals, where $m(\Delta)$ is so small that

$$\int_{C(\Delta)} f(x)\,dx$$

is less than the Harnack integral

$$\int_a^b f(x)\,dx$$

by less than an arbitrarily chosen positive number ζ. Let N be a positive number not less than the upper boundary of $f(x)$ in $C(\Delta)$, and let $f_N(x)$ be the function corresponding to $f(x)$ employed in de la Vallée Poussin's definition (§ 387). Let another set of intervals Δ', all interior to intervals of Δ, enclose all the points of infinite discontinuity of $f(x)$. The integral of $f(x)$ over $C(\Delta')$ lies between the integral over $C(\Delta)$ and Harnack's integral, and therefore differs from the last by less than ζ.

It follows that $$\int_{(\Delta-\Delta')} f(x)\,dx < \zeta,$$

and therefore $$\int_{(\Delta-\Delta')} f_N(x)\,dx < \zeta.$$

From this we deduce that

$$\int_{(\Delta)} f_N(x)\,dx < \zeta + N m(\Delta');$$

and since this holds for an arbitrarily small value of $m(\Delta')$, N being fixed, we have

$$\int_{(\Delta)} f_N(x)\,dx \leq \zeta.$$

It now follows that $\displaystyle\int_a^b f_N(x)\,dx - \int_{C(\Delta)} f(x)\,dx \leq \zeta;$

and since ζ is arbitrarily small, N being sufficiently increased, it follows that

$$\int_a^b f_N(x)\,dx$$

has a definite limit, as $N \sim \infty$, and that this limit is Harnack's integral

$$\int_a^b f(x)\,dx.$$

It has thus been shewn that the existence of Harnack's integral implies that of the integral as defined by de la Vallée Poussin, the integrals having the same value. By § 387, it follows that the L-integral exists, and is the same in value as the Harnack integral.

Accordingly, an absolutely convergent integral of a function such that the points of discontinuity form a set of measure zero, and which exists according to either Lebesgue's definition, or that of Harnack, exists also according to the other definition, and has the same value for the two definitions. It has moreover been shewn that the condition that each interval of Δ must contain at least one point of G, the set of points of infinite discontinuity of the function, is unnecessary when the integral is absolutely convergent.

THE LEBESGUE-STIELTJES INTEGRAL

445. Let $f(x)$ be a measurable function, defined for the linear interval (a, b), and let $\phi(x)$ be a bounded monotone non-diminishing function defined in the same interval. Denoting $\phi(x)$ by ξ, we consider, as in § 252, the functional image $E^{(\xi)}$, on the ξ-segment, of a set of points $E^{(x)}$ on the x-segment, such that to a point x', at which $\phi(x')$ is discontinuous, there corresponds the whole interval $(\phi(x'-0), \phi(x'+0))$ of points on the ξ-segment. It was remarked in § 252 that, if $E^{(x)}$ is measurable on the x-segment, it is not necessarily the case that $E^{(\xi)}$ is measurable, but that, if $E^{(x)}$ is measurable (B), then $E^{(\xi)}$ is measurable (B).

Let $F(\xi)$ be the function defined, at every point of the ξ-segment, by the condition $F(\xi) = f(x)$, where x is the point that corresponds to ξ; if x' is a point of discontinuity of $\phi(x)$, the interval $(\phi(x'-0), \phi(x'+0))$ is an interval of invariability of $F(\xi)$, in which $F(\xi) = \phi(x')$.

Whenever the L-integral $\int_\alpha^\beta F(\xi)\,d\xi$ *exists, where* $\alpha = \phi(a)$, $\beta = \phi(b)$, *its value may be said to define that of the Lebesgue-Stieltjes integral, or LS-integral* $\int_a^b f(x)\,d\phi(x)$, *of the measurable function* $f(x)$, *with respect to the monotone function* $\phi(x)$.

In case $\int_a^b F(\xi)\,d\xi$ exists as an R-integral, the integral $\int_a^b f(x)\,d\phi(x)$ may be an RS-integral (see § 378) It has been shewn, in § 377, that the necessary and sufficient condition that the integral may be an RS-integral is that the variation of $\phi(x)$ over the set of points of discontinuity of $f(x)$ should be zero.

If the function $f(x)$ is measurable (B), the function $F(\xi)$ is measurable (B) on the ξ-segment, and Lebesgue's definition is applicable to define $\int_a^\beta F(\xi)\,d\xi$.

In case $\phi(x)$ is of bounded variation, and is therefore expressible as the difference of two monotone non-diminishing functions $\phi_1(x)$, $\phi_2(x)$, the LS-integral of $f(x)$ with respect to $\phi(x)$ may be defined to be the excess of the LS-integral of $f(x)$ with respect to $\phi_1(x)$ over its LS-integral with respect to $\phi_2(x)$, whenever these latter integrals exist.

Referring to the definition, in § 252, of the variation of the monotone function $\phi(x)$ over a set of points E in the x-segment, and denoting by e_n that set of points at which $c_n \leq f(x) < c_{n+1}$, we see that the LS-integral $\int_a^b f(x)\,d\phi(x)$ is defined as the limit* of $\sum_{-\infty}^{\infty} c_n V^{(e_n)}\phi(x)$; the set of numbers $\{c_n\}$ being such that $c_{n+1} - c_n$ is less than a fixed number ϵ, for all positive and negative values of n, and the limit being taken as $\epsilon \sim 0$; it being assumed that the variation $V^{(e_n)}\phi(x)$ exists for all the sets e_n, so that $F(\xi)$ is measurable in (α, β).

The definition of an LS-integral, over the x-segment, as an L-integral over the ξ-segment, may be employed to extend the properties of L-integrals to the case of LS-integrals.

We have, for example, the extension of the property of L-integrals, that, if $\{f_n(x)\}$ is a sequence of summable functions that converges to $f(x)$, and such that $|f_n(x)|$ is bounded for all values of n and x, then

$$\int_a^b f(x)\,dx = \lim_{n\sim\infty} \int_a^b f_n(x)\,dx.$$

The corresponding property of the LS-integral is that

$$\int_a^b f(x)\,d\phi(x) = \lim_{n\sim\infty} \int_a^b f_n(x)\,d\phi(x),$$

---

* See Hildebrandt, "On integrals related to, and extensions of, Lebesgue integrals," *Bull. Amer. Math. Soc.* (2), vol. XXIV (1918), p. 191.

where the bounded functions $\{f_n(x)\}$ are such that their LS-integrals with respect to $\phi(x)$ exist.

446. It has been pointed out by Hildebrandt (*loc. cit.*) that the definition of an integral due to W. H. Young, referred to in § 389, may be so extended as to give rise to a definition of integration of a summable function with respect to a monotone function, more general than the definition given above.

Let the interval (a, b) be divided into a finite, or enumerably infinite, set of parts $e_1, e_2, \ldots e_n, \ldots$ over each of which the function $\phi(x)$ is measurable, and such that, over each of the sets e_n, the function $f(x)$ has finite upper and lower boundaries U_n, L_n. Consider the sums

$$S_1 = \Sigma U_n V^{(e_n)} \phi(x),\quad S_2 = \Sigma L_n V^{(e_n)} \phi(x);$$

then the lower boundary of S_1, for all such modes of division of (a, b) into the sum of sets, is defined to be the upper integral $\overline{\int_a^b} f(x)\, d\phi(x)$, of $f(x)$ with respect to $\phi(x)$, over the interval (a, b). Similarly, the lower integral $\underline{\int_a^b} f(x)\, d\phi(x)$ is defined to be the upper boundary of S_2, for all such modes of division of the interval (a, b). When the upper and lower integrals have the same value, that value may be said to define the integral $\int_a^b f(x)\, d\phi(x)$, of $f(x)$ with respect to $\phi(x)$. In accordance with this definition any summable function $f(x)$ will have* an integral with respect to $\phi(x)$, the division of (a, b) into parts being restricted to be a division into sets that are all measurable (B). The definition of § 445 is such that all functions that are measurable (B) have integrals, finite, or infinite, with respect to the monotone function $\phi(x)$.

Lebesgue has given† two modes by which a Stieltjes integral can be reduced to an L-integral. A transformation of an L-integral into a Stieltjes integral has been given‡ by Van Vleck. If $e(y)$ denotes the measure of the set of points for which $L \leq f(x) < y$, where L is the lower boundary of $f(x)$ in (a, b), the L-integral $\int_a^b f(x)\, dx$ is equal to the Stieltjes integral $\int_L^U y\, de(y)$, where U is the upper boundary of $f(x)$ in (a, b). This includes the case in which one, or both, of the numbers U, L are infinite. It has been shewn by Bliss§ that an L-integral is reducible to an R-integral.

For, $\int_L^U y\, de(y) = U(b-a) - \int_L^U e(y)\, dy$; and the last integral is an R-integral, since $e(y)$ is a monotone function of y.

---

* See Hildebrandt, *loc. cit.* p. 191.

† *Comptes Rendus*, Paris, vol CL (1910), p. 86.

‡ *Trans. Amer. Math. Soc.* vol. XVIII (1917), p. 326.

§ *Bull. Amer. Math. Soc.* vol. XXIV (1918), p. 1.

447. From the theorem of § 440 we see that, if $\Phi(x) = \int_a^x \phi(x)\,dx$, where $\Phi(x)$ is a monotone function of x, in an interval (a, b), then

$$\int_a^x f(x)\,\phi(x)\,dx$$

is equal to $\int_\alpha^\xi F(\xi)\,d\xi$, provided $F(\xi)$ is summable in the interval (α, β) that corresponds to (a, b), when the transformation $\xi = \alpha + \int_a^x \phi(x)\,dx$ is employed, and $F(\xi) = f(x)$. It then follows from the definition in § 445 that

$$\int_a^x f(x)\,\phi(x)\,dx = \int_a^x f(x)\,d\Phi(x)$$

This may be extended to the case in which $\Phi(x)$ is the indefinite integral of any summable function $\phi(x)$. For $\phi(x)$ may be expressed as

$$\phi_1(x) - \phi_2(x),$$

where $\phi_1(x)$, $\phi_2(x)$ are non-negative summable functions; and thus

$$\Phi(x) = \int_a^x \phi_1(x)\,dx - \int_a^x \phi_2(x)\,dx$$
$$= \Phi_1(x) - \Phi_2(x).$$

In accordance with the definition in § 445, we now have

$$\int_a^x f(x)\,\phi(x)\,dx = \int_a^x f(x)\,d\Phi_1(x) - \int_a^x f(x)\,d\Phi_2(x)$$

or

$$\int_a^x f(x)\,\phi(x)\,dx = \int_a^x f(x)\,d\Phi(x),$$

where $\Phi(x)$ denotes the indefinite integral of the summable function $\phi(x)$.

Let $f(x)$ be monotone and bounded in the interval $(a. b)$, then

$$\int_a^b f(x)\,d\Phi(x)$$

is an *RS*-integral, since the variation of $\Phi(x)$ over the set of points of discontinuity of $f(x)$ is zero, because $\Phi(x)$ is an indefinite integral.

We then have (see § 376)

$$\int_a^b f(x)\,d\Phi(x) = \Big[f(x)\,\Phi(x)\Big]_a^b - \int_a^b \Phi(x)\,df(x),$$

and thus the theorem for integration by parts may be written in the form

$$\int_a^b f(x)\,\phi(x)\,dx = \Big[f(x)\,\Phi(x)\Big]_a^b - \int_a^b \Phi(x)\,df(x).$$

Moreover, since $f(x)$ is monotone,

$$\int_a^b \Phi(x)\,df(x) = \Phi(\mu)\,\{f(b) - f(a)\},$$

where μ is some point in the interval (a, b).

Therefore $\int_a^b f(x)\,\phi(x)\,dx = f(b)\,\Phi(b) - \Phi(\mu)\{f(b) - f(a)\}$

$$= f(a)\int_a^\mu \phi(x)\,dx + f(b)\int_\mu^b \phi(x)\,dx.$$

A proof of the second mean value theorem* has thus been obtained by employing the method of integration by parts, as applied to the *RS*-integral. The more general form of the theorem, and Bonnet's form, can be deduced, as in § 422.

THE *RS*-INTEGRAL FOR FUNCTIONS OF TWO VARIABLES

448. Let $f(x^{(1)}, x^{(2)})$, $g(x^{(1)}, x^{(2)})$ both be functions of bounded variation, in accordance with the definition of § 254. Let the functions be such that each of them has an *RS*-integral with respect to the other†, taken over the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. Then these integrals satisfy the following relation:

$$\text{(A).}\quad \int_{(a^{(1)},a^{(2)})}^{(b^{(1)},b^{(2)})} g(x^{(1)}, x^{(2)})\,df(x^{(1)}, x^{(2)}) = \Big[f(x^{(1)}, x^{(2)})\,g(x^{(1)}, x^{(2)})\Big]_{(a^{(1)},a^{(2)})}^{(b^{(1)},b^{(2)})}$$
$$- \int_{a^{(1)}}^{b^{(1)}} \Big[f(x^{(1)}, x^{(2)})\,dg(x^{(1)}, x^{(2)})\Big]_{a^{(2)}}^{b^{(2)}}$$
$$- \int_{a^{(2)}}^{b^{(2)}} \Big[f(x^{(1)}, x^{(2)})\,dg(x^{(1)}, x^{(2)})\Big]_{a^{(1)}}^{b^{(1)}}$$
$$+ \int_{(a^{(1)},a^{(2)})}^{(b^{(1)},b^{(2)})} f(x^{(1)}, x^{(2)})\,dg(x^{(1)}, x^{(2)}).$$

where $\quad \Big[f(x^{(1)}, x^{(2)})\,dg(x^{(1)}, x^{(2)})\Big]_{a^{(1)}}^{b^{(1)}}$

denotes $\quad f(b^{(1)}, x^{(2)})\,dg(b^{(1)}, x^{(2)}) - f(a^{(1)}, x^{(2)})\,dg(a^{(1)}, x^{(2)})$,

and $\quad \Big[f(x^{(1)}, x^{(2)})\,g(x^{(1)}, x^{(2)})\Big]_{(a^{(1)},a^{(2)})}^{(b^{(1)},b^{(2)})}$

denotes

$$f(b^{(1)}, b^{(2)})\,g(b^{(1)}, b^{(2)}) - f(b^{(1)}, a^{(2)})\,g(b^{(1)}, a^{(2)}) - f(a^{(1)}, b^{(2)})\,g(a^{(1)}, b^{(2)})$$
$$+ f(a^{(1)}, a^{(2)})\,g(a^{(1)}, a^{(2)}).$$

In order to prove this formula, let the rectangle $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$ be divided into $(m-1)(n-1)$ partial rectangles, and let $P_{i,j}$ denote a corner of one of these partial rectangles, where i has the values $1, 2, \ldots m-1$, and j has the values $1, 2, \ldots n-1$. We have then the identity

$$\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\big[g_{i,j}(f_{i,j} - f_{i+1,j} - f_{i,j+1} + f_{i+1,j+1})$$
$$- f_{i+1,j+1}(g_{i,j} - g_{i+1,j} - g_{i,j+1} + g_{i+1,j+1})\big]$$
$$= \sum_{i=1}^{m-1} f_{i+1,1}(g_{i+1,1} - g_{i,1}) + \sum_{j=1}^{n-1} f_{1,j+1}(g_{1,j+1} - g_{1,j})$$
$$- \sum_{i=1}^{m-1} f_{i+1,n}(g_{i+1,n} - g_{i,n}) - \sum_{j=1}^{n-1} f_{m,j+1}(g_{m,j+1} - g_{m,j})$$
$$+ [f_{1,1}g_{1,1} - f_{1,n}g_{1,n} - f_{m,1}g_{m,1} + f_{m,n}g_{m,n}];$$

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p. 280.

† This can be shewn to be the case if one of the functions is continuous.

where $f_{i,j}$, $g_{i,j}$ denote the values of $f(x^{(1)}, x^{(2)})$, $g(x^{(1)}, x^{(2)})$ at the point $P_{i,j}$.

If we assume that the functions $f(x^{(1)}, x^{(2)})$, $g(x^{(1)}, x^{(2)})$ are both of bounded variation in accordance with the definition of § 254, and that each of them has an *RS*-integral with respect to the other, as the numbers m and n are indefinitely increased, and the rectangular sub-division is such that the maximum diagonal of the partial rectangles converges to zero, the integrals in the formula (A) will be given as the limits of the summations in the above identity.

The formula (A) may be reduced to a different form which is one of four different forms.

It can easily be verified that

$$\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, b^{(2)})\, dg(x^{(1)}, x^{(2)}) = \int_{a^{(1)}}^{b^{(1)}} f(x^{(1)}, b^{(2)})\, d\Big[g(x^{(1)}, x^{(2)})\Big]_{a^{(2)}}^{b^{(2)}}.$$

where $\Big[g(x^{(1)}, x^{(2)})\Big]_{a^{(2)}}^{b^{(2)}}$ denotes $g(x^{(1)}, b^{(2)}) - g(x^{(1)}, a^{(2)})$.

Employing this result, we have

$$\begin{aligned}
&\int_{a^{(1)}}^{b^{(1)}} \Big[\int f(x^{(1)}, x^{(2)})\, dg(x^{(1)}, x^{(2)})\Big]_{a^{(2)}}^{b^{(2)}} \\
&= \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, b^{(2)})\, dg(x^{(1)}, x^{(2)}) \\
&\qquad + \int_{a^{(1)}}^{b^{(1)}} [f(x^{(1)}, b^{(2)}) - f(x^{(1)}, a^{(2)})]\, dg(x^{(1)}, a^{(2)}) \\
&= \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f(x^{(1)}, b^{(2)})\, dg(x^{(1)}, x^{(2)}) - \int_{a^{(1)}}^{b^{(1)}} \Big[f(x^{(1)}, x^{(2)})\Big]_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \\
&\qquad + [f(b^{(1)}, b^{(2)}) - f(b^{(1)}, a^{(2)})]\, [g(b^{(1)}, a^{(2)}) - g(a^{(1)}, a^{(2)})].
\end{aligned}$$

A similar expression for

$$\int_{a^{(2)}}^{b^{(2)}} \Big[\int f(x^{(1)}, x^{(2)})\, dg(x^{(1)}, x^{(2)})\Big]_{a^{(1)}}^{b^{(1)}}$$

can be obtained. On substituting these results in (A), we obtain the formula

$$\begin{aligned}
(\mathrm{B}_1).\quad &\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} g(x^{(1)}, x^{(2)})\, df(x^{(1)}, x^{(2)}) \\
&= g(a^{(1)}, a^{(2)}) \Big[f(x^{(1)}, x^{(2)})\Big]_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \\
&\qquad + \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \Big[f(x^{(1)}, x^{(2)})\Big]_{(x^{(1)}, x^{(2)})}^{(b^{(1)}, b^{(2)})} dg(x^{(1)}, x^{(2)}) \\
&\qquad + \int_{a^{(1)}}^{b^{(1)}} \Big[f(x^{(1)}, x^{(2)})\Big]_{(x^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} dg(x^{(1)}, a^{(2)}) \\
&\qquad + \int_{a^{(2)}}^{b^{(2)}} \Big[f(x^{(1)}, x^{(2)})\Big]_{(a^{(1)}, x^{(2)})}^{(b^{(1)}, b^{(2)})} dg(a^{(1)}, x^{(2)}).
\end{aligned}$$

In a similar manner, or by interchange of the corners of the cell, we obtain the formulae (B_2), (B_3), (B_1). For example, the formula (B_2) is

$$\begin{aligned}\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} g\,(x^{(1)}, x^{(2)})\, df\,(x^{(1)}, x^{(2)}) - g\,(a^{(2)}, b^{(2)}) \Big| f\,(x^{(1)}, x^{(2)}) \Big|_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \\ + \int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} \Big[f\,(x^{(1)}, x^{(2)}) \Big]_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, x^{(2)})} dg\,(x^{(1)}, x^{(2)}) \\ - \int_{a^{(1)}}^{b^{(1)}} \Big[f\,(x^{(1)}, x^{(2)}) \Big]_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, b^{(2)})} dg\,(x^{(1)}, b^{(2)}) \\ - \int_{a^{(2)}}^{b^{(2)}} \Big[f\,(x^{(1)}, x^{(2)}) \Big]_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, x^{(2)})} dg\,(a^{(2)}, x^{(2)}).\end{aligned}$$

The formulae for functions of three variables have been given* by W. H. Young.

448^1. Let $f\,(x^{(1)}, x^{(2)})$ be a function of bounded variation, in accordance with the definition of § 254, and let $G\,(x^{(1)}, x^{(2)})$ denote the indefinite integral $\int_{(a^{(1)}, a^{(2)})}^{(x^{(1)}, x^{(2)})} g\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)})$, where $g\,(x^{(1)}, x^{(2)})$ is any function which is summable in the cell $(a^{(1)}, a^{(2)}; b^{(1)}, b^{(2)})$. It will then be shewn that *the L-integral* $\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f\,(x^{(1)}, x^{(2)})\, g\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)})$ *is equal to* $\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f\,(x^{(1)}, x^{(2)})\, dG\,(x^{(1)}, x^{(2)})$, *which exists since* $G\,(x^{(1)}, x^{(2)})$ *is absolutely continuous*. For its variation over the set of points of discontinuity of $f\,(x^{(1)}, x^{(2)})$, of measure zero (§ 308), is zero (see § 381).

To prove the theorem, let it be assumed, in the first instance, that $g\,(x^{(1)}, x^{(2)})$ is a non-negative function, so that $G\,(x^{(1)}, x^{(2)})$ is quasi-monotone.

We see that $\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f\,(x^{(1)}, x^{(2)})\, g\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)})$

or
$$\sum_{r=1}^{r=m} \int_{(\alpha_r^{(1)}, \alpha_r^{(2)})}^{(\beta_r^{(1)}, \beta_r^{(2)})} f\,(x^{(1)}, x^{(2)})\, g\,(x^{(1)}, x^{(2)})\, d\,(x^{(1)}, x^{(2)}),$$

in accordance with the notation of § 381, lies between

$$\sum_{r=1}^{r=m} U\,(\delta_r)\, \Delta_{\delta_r} G\,(x^{(1)}, x^{(2)}) \text{ and } \sum_{r=1}^{r=m} L\,(\delta_r)\, \Delta_{\delta_r} G\,(x^{(1)}, x^{(2)}),$$

both of which converge to $\int_{(a^{(1)}, a^{(2)})}^{(b^{(1)}, b^{(2)})} f\,(x^{(1)}, x^{(2)})\, dG\,(x^{(1)}, x^{(2)})$, as m is indefinitely increased. Thus the theorem holds when $g\,(x^{(1)}, x^{(2)})$ is non-negative. Any summable function $g\,(x^{(1)}, x^{(2)})$ can be expressed as the difference of two non-negative functions $g_1\,(x^{(1)}, x^{(2)})$ and $g_2\,(x^{(1)}, x^{(2)})$, to each of which the theorem applies; it then follows that the theorem holds generally.

---

* See W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p. 281.

HELLINGER'S INTEGRALS

449. In connection with the theory of the reduction of a quadratic form involving an infinite number of variables to a canonical quadratic form, certain limits were introduced by Hellinger*, for the representation of which he used the notation of integration. A limit, of this species. may accordingly be spoken of as a Hellinger integral. It was shewn by Hahn that every Hellinger integral can be expressed as an L-integral.

Let $g(y)$ be a continuous non-decreasing monotone function, defined for the linear interval (α, β); and let $f(y)$ be a summable function, defined for the same interval, and such that it is constant in any interval. contained in (α, β), in which $g(y)$ is constant.

Let a net be fitted on to the interval (α, β), and denote its meshes by (y_0, y_1), (y_1, y_2), ... (y_{m-1}, y_m). where $y_0 = \alpha$, $y_m = \beta$.

If the sum $\sum_{r=1}^{r=m} \frac{\{f(y_r) - f(y_{r-1})\}^2}{g(y_r) - g(y_{r-1})}$ *has a finite upper boundary, when all possible nets are taken into account, that upper boundary is denoted by* $\int_\alpha^\beta \frac{[df(y)]^2}{dg(y)}$, *and is called a Hellinger integral. In the above sum, any term in which the denominator, and consequently the numerator, vanishes, is omitted.*

Let the new variable x be defined for the interval (a, b), where $a = g(\alpha)$, $b = g(\beta)$, by the definition $x = g(y)$.

The sum employed in the definition of the Hellinger integral becomes $\sum_{r=1}^{r=m} \frac{\{F(x_r) - F(x_{r-1})\}^2}{x_r - x_{r-1}}$, where x_r corresponds to y_r, and $F(x)$, or $F\{g(y)\}$, is identical with $f(y)$.

If $f_1(y)$, $f_2(y)$ *are two summable functions, defined in* (α, β), *and each is constant in any interval in which* $g(y)$ *is constant, then if the sum*

$$\sum_{r=1}^{r=m} \frac{\{f_1(y_r) - f_1(y_{r-1})\}\{f_2(y_r) - f_2(y_{r-1})\}}{g(y_r) - g(y_{r-1})}$$

has a definite limit, the same for all systems of nets fitted on to (α, β), *this limit is denoted by* $\int_\alpha^\beta \frac{df_1(y)\,df_2(y)}{dg(y)}$, *and is also called a Hellinger integral.*

The sum here employed is equivalent to

$$\sum_{r=1}^{r=m} \frac{\{F_1(x_r) - F_1(x_{r-1})\}\{F_2(x_r) - F_2(x_{r-1})\}}{x_r - x_{r-1}},$$

where $F_1\{g(y)\} \equiv f_1(y)$, and $F_2\{g(y)\} \equiv f_2(y)$.

---

* See the dissertation "Die orthogonal invarianten quadratischen Formen von unendlich vielen Veranderlichen," Gottingen, 1907; also a memoir in *Crelle's Journal*, vol. CXXXVI (1909), p. 210.

450. The following theorem will be established:

If $\phi(x)$ *denote a function, bounded, or unbounded, in* (a, b), *such that* $\{\phi(x)\}^2$ *is summable in the interval, then*

$$\frac{1}{\delta_1}\left\{\int_{(\delta_1)}\phi(x)\,dx\right\}^2+\frac{1}{\delta_2}\left\{\int_{(\delta_2)}\phi(x)\,dx\right\}^2+\ldots+\frac{1}{\delta_m}\left\{\int_{(\delta_m)}\phi(x)\,dx\right\}^2,$$

where $\delta_1, \delta_2, \ldots \delta_m$ *are the intervals of a net fitted on to* (a, b), *converges to* $\int_a^b\{\phi(x)\}^2dx$, *for any system of nets. If* $\{\phi(x)\}^2$ *is not summable, the above sum diverges, for a system of nets.*

In case $\phi(x)$ is integrable (R), the truth of the theorem is obvious; for $\int_{(\delta_r)}\phi(x)\,dx$ is equal to δ_r multiplied by a number lying between the upper and lower boundaries of $\phi(x)$ in δ_r; and thus the expression reduces to the Riemann sum employed in the definition of $\int_a^b\{\phi(x)\}^2dx$. Thus the theorem holds, in particular, for a continuous function $\phi(x)$.

Next, let $\phi(x)$ be any function that is non-negative, summable, and bounded, in (a, b). A continuous function $f(x)$, also non-negative, can be so determined that

$$\int_a^b|\{\phi(x)\}^2-\{f(x)\}^2|\,dx<\epsilon,$$

and consequently that $\int_a^b\{\phi(x)-f(x)\}^2dx<\epsilon$, (see § 433); where ϵ is an arbitrarily prescribed positive number.

The sum $\sum\limits_{r=1}^{r=m}\frac{1}{\delta_r}\left\{\int_{(\delta_r)}\phi(x)\,dx\right\}^2$ may be expressed by $S_1+S_2+S_3$, where

$$S_1=\sum_{r=1}^{r=m}\frac{1}{\delta_r}\left\{\int_{(\delta_r)}\{\phi(x)-f(x)\}\,dx\right\}^2,\quad S_2=\sum_{r=1}^{r=m}\frac{1}{\delta_r}\left\{\int_{(\delta_r)}f(x)\,dx\right\}^2,$$

and $$S_3=2\sum_{r=1}^{r=m}\frac{1}{\delta_r}\int_{(\delta_r)}\{\phi(x)-f(x)\}\,dx\int_{(\delta_r)}f(x)\,dx.$$

By using the Schwarzian inequality (§ 396), we have

$$\left\{\int_{(\delta_r)}\{\phi(x)-f(x)\}\,dx\right\}^2\leq\delta_r\int_{(\delta_r)}\{\phi(x)-f(x)\}^2dx;$$

and hence we find that $S_1<\epsilon$. Also, since $f(x)$ is continuous, we have, for all nets of sufficiently large order, in any system of nets,

$$\int_a^b\{f(x)\}^2dx>S_2>\int_a^b\{f(x)\}^2dx-\epsilon.$$

We also have

$$|S_3|<2U\int_a^b|\phi(x)-f(x)|\,dx<2U\left[(b-a)\int_a^b\{\phi(x)-f(x)\}^2dx\right]^{\frac{1}{2}};$$

and thus $|S_3|<2U(b-a)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}$; where U denotes the upper boundary of $\phi(x)$; since the function $f(x)$ may be so chosen that its upper boundary does not exceed that of $\phi(x)$.

Since ϵ is arbitrary, S_1 and S_3 are arbitrarily small; and S_2 differs from $\int_a^b \{\phi(x)\}^2\, dx$ by less than 2ϵ. It now follows that

$$\sum_{r=1}^{r=m} \frac{1}{\delta_r}\left\{\int_{(\delta_r)} \phi(x)\, dx\right\}^2$$

converges to $\int_a^b \{\phi(x)\}^2\, dx$, as the summation is taken successively over the nets of any system of nets.

Next, let $\phi(x)$ be unbounded, but everywhere $\geqq 0$; and let $\phi_{N_s}(x)$ have the same value as $\phi(x)$, at a point where $\phi(x) \leqq N_s$, and have the value N_s, at a point where $\phi(x) > N_s$.

Denoting by a_{ms} the number $\sum_{r=1}^{r=m} \frac{1}{\delta_r}\left\{\int_{(\delta_r)} \phi_{N_s}(x)\, dx\right\}^2$; as s has the values in an increasing sequence such that $N_s \sim \infty$, and m increases indefinitely as we proceed through the nets of a given system, the numbers a_{ms} form a monotone double sequence. For a sub-division of any net increases the value of a_{ms}, as is easily seen from the inequality $\frac{(p_1+p_2)^2}{k_1+k_2} \leqq \frac{p_1^2}{k_1} + \frac{p_2^2}{k_2}$. If either of the repeated limits $\lim_{m\sim\infty}\lim_{s\sim\infty} a_{ms}$, $\lim_{s\sim\infty}\lim_{m\sim\infty} a_{ms}$ exists, then also the other exists, and the two have the same value (see § 388).

Now $\lim_{m\sim\infty} a_{ms} = \int_a^b \{\phi_{N_s}(x)\}^2\, dx$, and $\lim_{s\sim\infty}\lim_{m\sim\infty} a_{ms} = \int_a^b \{\phi(x)\}^2\, dx$,

hence
$$\lim_{m\sim\infty}\lim_{s\sim\infty} a_{ms} = \int_a^b \{\phi(x)\}^2\, dx,$$

and thus the limit, as we proceed through a system of nets, of

$$\sum_{r=1}^{r=m} \frac{1}{\delta_r}\left\{\int_{(\delta_r)} \phi(x)\, dx\right\}^2, \text{ is } \int_a^b \{\phi(x)\}^2\, dx.$$

Lastly, let $\phi(x)$ be any summable function whose square is summable, and which may have both positive and negative values It may be expressed as the difference of two summable functions $\phi_1(x)$, $\phi_2(x)$, each of which is $\geqq 0$, for all values of x.

Since $\Sigma \frac{1}{\delta_r}\left[\int_{(\delta_r)} \{\phi_1(x) + \phi_2(x)\}^2\, dx\right]^2$ converges to $\int_a^b \{\phi_1(x) + \phi_2(x)\}^2\, dx$, that is to $\int_a^b \{\phi_1(x)\}^2\, dx + \int_a^b \{\phi_2(x)\}^2\, dx + 2\int_a^b \phi_1(x)\,\phi_2(x)\, dx$; and also

$$\Sigma \frac{1}{\delta_r}\left\{\int_{(\delta_r)} \phi_1(x)\, dx\right\}^2 \text{ converges to } \int_a^b \{\phi_1(x)\}^2\, dx,$$

and
$$\Sigma \frac{1}{\delta_r}\left\{\int_{(\delta_r)} \phi_2(x)\, dx\right\}^2 \text{ converges to } \int_a^b \{\phi_2(x)\}^2\, dx,$$

it follows that $\Sigma \frac{1}{\delta_r}\int_{(\delta_r)} \phi_1(x)\, dx \int_{(\delta_r)} \phi_2(x)\, dx$ converges to $\int_a^b \phi_1(x)\,\phi_2(x)\, dx$.

It now appears that

$$\Sigma \frac{1}{\delta_r}\left\{\int_{(\delta_r)}\{\phi_1(x)-\phi_2(x)\}^2\,dx \right. \text{ converges to } \int_a^b \{\phi_1(x)-\phi_2(x)\}^2\,dx,$$

and thus the theorem is established generally for every function $\phi(x)$ whose square is summable.

The following theorem, which includes the theorem established above, may now be stated:

If $\phi_1(x)$, $\phi_2(x)$ be any functions whose squares are summable in the interval (a, b), the sum

$$\sum_{r=1}^{r=m}\frac{1}{\delta_r}\int_{(\delta_r)}\phi_1(x)\,dx\int_{(\delta_r)}\phi_2(x)\,dx \text{ converges to } \int_a^b \phi_1(x)\,\phi_2(x)\,dx,$$

as the summation is taken successively over the nets of any system of nets fitted on to (a, b).

To prove this theorem, we observe that $\phi_1(x)$, $\phi_2(x)$ may be replaced by $\phi_1^+(x)-\phi_1^-(x)$, $\phi_2^+(x)-\phi_2^-(x)$, respectively, where the four functions in these expressions are all non-negative. As the theorem holds for each of the integrals

$$\int_a^b \phi_1^+(x)\,\phi_2^+(x)\,dx,\quad \int_a^b \phi_1^+(x)\,\phi_2^-(x)\,dx,$$
$$\int_a^b \phi_1^-(x)\,\phi_2^+(x)\,dx,\quad \int_a^b \phi_1^-(x)\,\phi_2^-(x)\,dx,$$

it clearly holds also for

$$\int_a^b \{\phi_1^+(x)-\phi_1^-(x)\}\{\phi_2^+(x)-\phi_2^-(x)\}\,dx,$$

that is for $$\int_a^b \phi_1(x)\,\phi_2(x)\,dx.$$

451. Let $x = g(y)$, $a = g(\alpha)$, $b = g(\beta)$, where $g(y)$ is a continuous, non-diminishing function of y, in the interval (α, β); also let $f(y) = F\{g(y)\}$, where $F(x) = \int_a^x \phi(x)\,dx$.

The sum employed in the theorem of § 450 then becomes

$$\sum_{r=1}^{r=m}\frac{\{f(y_r)-f(y_{r-1})\}^2}{g(y_r)-g(y_{r-1})},$$

where the net $(y_0, y_1), (y_1, y_2), \ldots (y_{m-1}, y_m)$ corresponds to the net $(x_0, x_1), (x_1, x_2), \ldots (x_{m-1}, x_m)$.

It will now be shewn that:

The necessary and sufficient condition that $\int_\alpha^\beta \frac{[df(y)]^2}{dg(y)}$ should exist is that, if $x = g(y)$, the function $F(x) \equiv f(y)$ should be the indefinite integral of a function $\phi(x)$ whose square is summable in the interval (a, b), of x.*

---

* This theorem was given by Hahn, see *Monatshefte für Math. u. Physik*, vol. XXIII (1912), p. 172. Another proof was given by Hobson, *Proc. Lond. Math. Soc.* (2), vol. XVIII (1918), p. 258.

Moreover $\int_{\alpha}^{\beta} \frac{[df(y)]^2}{dg(y)} = \int_{a}^{b} \{\phi(x)\}^2 dx$; *the Hellinger integral being thus expressed as an L-integral.*

To establish the sufficiency of the condition, let it be assumed that $F(x) = \int_{a}^{x} \phi(x)\, dx$, where $\{\phi(x)\}^2$ is summable.

We have then $F(x_r) - F(x_{r-1}) = \int_{x_{r-1}}^{x_r} \phi(x)\, dx$, and thence we see that

$$\{F(x_r) - F(x_{r-1})\}^2 \leq (x_r - x_{r-1}) \int_{x_{r-1}}^{x_r} \{\phi(x)\}^2 dx,$$

and therefore $$\sum_{r=1}^{r=m} \frac{\{F(x_r) - F(x_{r-1})\}^2}{x_r - x_{r-1}} \leq \int_{a}^{b} \{\phi(x)\}^2 dx,$$

for any net fitted on to (a, b). It follows that the limit of $\int_{\alpha}^{\beta} \frac{[df(y)]^2}{dg(y)}$ exists.

Next, let the existence of the Hellinger integral be assumed. It is evident, from the definition, that $\int_{\alpha}^{y} \frac{[df(y)]^2}{dg(y)}$ exists, for any value of y in the interval (α, β), and that it is a monotone increasing function of y, say $\psi(y)$.

We have then, in any interval of y, contained in (α, β),

$$\frac{\{\Delta f(y)\}^2}{\Delta g(y)} \leq \Delta\psi(y);$$

and thus

$$|F(x_r) - F(x_{r-1})| \leq [(x_r - x_{r-1})\{\chi(x_r) - \chi(x_{r-1})\}]^{\frac{1}{2}}, \text{ where } \psi(y) = \chi(x).$$

From this, we have

$$\sum_{r=1}^{r=m} |F(x_r) - F(x_{r-1})| \leq \sum_{r=1}^{r=m} [(x_r - x_{r-1})\{\chi(x_r) - \chi(x_{r-1})\}]^{\frac{1}{2}}$$
$$\leq [(b-a)\{\chi(b) - \chi(a)\}]^{\frac{1}{2}}.$$

and thus the function $F(x)$ is of bounded variation in (a, b): from which it follows that $F'(x)$ exists for almost all the values of x. In order to shew that $F(x)$ is the indefinite integral of $F'(x)$, it is sufficient to shew that the variation of $F(x)$ over any set of points of measure zero vanishes (see § 407). Such a set can be enclosed in the intervals (δ) of a non-overlapping set of which the measure is arbitrarily small. The variation of $F(x)$ over this set of intervals is, as above, $\leq [\Sigma\delta\, \{\chi(b) - \chi(a)\}]^{\frac{1}{2}}$, where $\Sigma\delta$ is arbitrarily small; hence the variation is zero. Denoting $F'(x)$, at every point at which it exists, by $\phi(x)$, we may suppose $\phi(x)$ to have the value zero at all points at which $F'(x)$ does not exist. The function $\phi(x)$ being summable in (a, b), the theorem of § 450 is applicable, and it follows that $\int_{a}^{b} \{\phi(x)\}^2 dx$ exists, and is the limit of the sum

$$\sum_{r=1}^{r=m} \frac{[F(x_r) - F(x_{r-1})]^2}{x_r - x_{r-1}},$$

for any system of nets fitted on to (a, b). Thus we have

$$\int_\alpha^\beta \frac{[df(y)]^2}{dg(y)} = \int_a^b \{\phi(x)\}^2\, dx = \int_a^b \{F'(x)\}^2\, dx,$$

and accordingly the theorem has been established.

It is sufficient for the existence of the Hellinger integral

$$\int_\alpha^\beta \frac{df_1(y)\, df_2(y)}{dg(y)}$$

that the functions $F_1(x)$, $F_2(x)$ should be the indefinite integrals of two functions $\phi_1(x)$, $\phi_2(x)$ whose squares are summable in (a, b); where $F_1(x)$, $F_2(x)$ are the functions which are equivalent to $f_1(y)$, $f_2(y)$, and $x = g(y)$. In accordance with the second theorem of § 449, we then see that

$$\int_a^b \phi_1(x)\, \phi_2(x)\, dx = \int_\alpha^\beta \frac{df_1(y)\, df_2(y)}{dg(y)}.$$

452. The theorems of § 450 can be generalized so as to relate to a function $\phi(x)$, such that $\{\phi(x)\}^{1+\lambda}$ is summable in the interval (a, b), where λ is a positive number. The proofs of the theorems in the extended form are essentially similar to those of the original theorems. Thus we obtain* the theorems:

If $\phi(x)$ be a function that is non-negative in the interval (a, b), and such that $\{\phi(x)\}^{1+\lambda}$ is summable in (a, b), where λ denotes some positive number, then the sum

$$\sum_{r=1}^{r=m} \frac{1}{\delta_r^{\lambda}} \left\{\int_{(\delta_r)} \phi(x)\, dx\right\}^{1+\lambda}$$

converges to $\int_a^b \{\phi(x)\}^{1+\lambda}\, dx$, as the sum is taken successively over the nets of any system of nets fitted on to (a, b). In case λ is a positive integer, $\phi(x)$ may be any summable function, not necessarily of one sign in (a, b).

Hellinger's definition may be extended to apply to the integral

$$\int_\alpha^\beta \frac{[df(y)]^{1+\lambda}}{[dg(y)]^{\lambda}},$$

defined as the limit, when it exists, of the sum

$$\sum_{r=1}^{r=m} \frac{[f(y_r) - f(y_{r-1})]^{1+\lambda}}{[g(y_r) - g(y_{r-1})]^{\lambda}}.$$

In this case:

The necessary and sufficient condition that $\int_\alpha^\beta \frac{[df(y)]^{1+\lambda}}{[dg(y)]^{\lambda}}$ should exist is that the function $F(x)$ $(= f(y))$ should be the indefinite integral of a function $\phi(x)$ such that $[\phi(x)]^{1+\lambda}$ is summable in (a, b). Moreover

$$\int_\alpha^\beta \frac{[df(y)]^{\lambda+1}}{[dg(y)]^{\lambda}} = \int_a^b \{\phi(x)\}^{1+\lambda}\, dx.$$

Other generalizations of the Hellinger integral have been made by Radon† and by E. H. Moore‡.

---

* See Hobson, *loc. cit.* p. 263 See also F. Riesz, *Math. Annalen*, vol. LXIX (1910), p. 462.

† See the memoir "Absolut additive Mengenfunktionen," *Wiener Sitzungsber.* vol. CXXII$_{2a}$ (1913), p. 1351.

‡ See Hildebrandt, *Bull. Amer. Math. Soc.* (2), vol. XXIV (1918), p. 198.

CHAPTER VIII

NON-ABSOLUTELY CONVERGENT INTEGRALS

453. The definitions of Harnack and Lebesgue have been shewn (§ 444) to lead to the same values in the case of absolutely convergent integrals when the points of infinite discontinuity of the function form a set of measure zero. We proceed to consider the integral of a function $f(x)$ over a linear interval, in accordance with Harnack's definition, when that integral is non-absolutely convergent, that is, when $|f(x)|$ is not integrable over (a, b) in accordance with Harnack's definition. A point of (a, b) which is such that $f(x)$ is not summable in any interval which contains the point within it, or at an end-point, may be called a *point of non-summability*, or a *Harnack point*, for the function $f(x)$. It is clear that the points of non-summability of $f(x)$ form a closed set H, for if P be a limiting point of H, any interval that contains P contains points of the set, and consequently $f(x)$ is not summable in the interval. The points of H are points of infinite discontinuity of $f(x)$, but H is not necessarily identical with the set G, employed in § 443, of all the points of infinite discontinuity of $f(x)$. In any interval interior to a contiguous interval of H the function $f(x)$ is summable, though it is not necessarily bounded. The definition of § 443 will be modified, by employing the set H instead of G; so that the L-integrals employed in the definition are not necessarily L-integrals of a bounded function, those points of infinite discontinuity of $f(x)$ which are not points of non-summability being interior to the contiguous intervals of H. Thus the definition takes the following form:

If there be a non-dense closed set H, of content zero, of points of non-summability of $f(x)$ in (a, b), and H be enclosed within intervals of a finite set Δ, such that each interval of Δ contains at least one point of H, then if the L-integral $\int_{C(\Delta)} f(x)\,dx$ exists, and is such that, corresponding to an arbitrarily chosen number ϵ, a number ζ_ϵ, which converges with ϵ to zero, exists such that

$$\left| \int_{C(\Delta)} f(x)\,dx - \int_{C(\Delta')} f(x)\,dx \right| < \epsilon,$$

for all such pairs of sets of intervals Δ, Δ', provided $m(\Delta) < \zeta_\epsilon$, $m(\Delta') < \zeta_\epsilon$, then the limit of $\int_{C(\Delta)} f(x)\,dx$, as $m(\Delta) \sim 0$, defines a number which is denoted

by $\int_a^b f(x)\,dx$, *and is called the Harnack-Lebesgue, or HL-integral, of* $f(x)$, *in* (a, b).

The HL-integrals of functions contain as a sub-class the Harnack-Riemann, or HR-integrals, in which the integral of the function over any interval interior to a contiguous interval of H is an R-integral.

It will appear that the condition that each interval of Δ must contain a point of H is an indispensable condition. It has been shewn (§ 444) that, in the case of absolutely convergent integrals, the corresponding condition relating to G is unnecessary. When this condition is a necessary one, the integral has been termed* by E. H. Moore a narrow integral. In the contrary case the integral is called a broad integral; and it will appear that a broad integral is necessarily absolutely convergent, so that the set H does not exist. The L-integral in the definition may be quite general.

It will now be shewn that, if $f(x)$ have an HL-integral in (a, b), it has an HL-integral in any interval (a', b') contained in (a, b).

Taking two sets of intervals Δ, Δ', as in the definition given above, we have, for every value of ϵ,

$$\left|\int_a^b f_\Delta(x)\,dx - \int_a^b f_{\Delta'}(x)\,dx\right| < \epsilon,$$

provided $m(\Delta) < \zeta_\epsilon$, $m(\Delta') < \zeta_\epsilon$; where $f_\Delta(x) = f(x)$, in the intervals $C(\Delta)$, and $f_\Delta(x) = 0$, in Δ: with a corresponding definition for $f_{\Delta'}(x)$. Assuming that this condition is satisfied for every value of ϵ, it will be shewn that

$$\left|\int_{a'}^{b'} f_\Delta(x)\,dx - \int_{a'}^{b'} f_{\Delta'}(x)\,dx\right| < \epsilon,$$

provided Δ, Δ' satisfy the more stringent conditions

$$m(\Delta) < \tfrac{1}{2}\zeta_\epsilon, \quad m(\Delta') < \tfrac{1}{2}\zeta_\epsilon.$$

Let it be assumed that, if possible, Δ and Δ' can be determined so as to satisfy these last conditions, and so that

$$\left|\int_{a'}^{b'} f_\Delta(x)\,dx - \int_{a'}^{b'} f_{\Delta'}(x)\,dx\right| \geq \epsilon.$$

It will then be shewn that finite sets of intervals $\bar{\Delta}, \bar{\Delta}'$ can be determined which satisfy the condition that each interval of either set contains at least one point of H, and such that

$$m(\bar{\Delta}) < \zeta_\epsilon, \quad m(\bar{\Delta}') < \zeta_\epsilon, \quad \left|\int_a^b f_{\bar{\Delta}}(x)\,dx - \int_a^b f_{\bar{\Delta}'}(x)\,dx\right| \geq \epsilon;$$

and since this is contrary to the hypothesis made above, the impossibility of the assumption will have been demonstrated.

To define $\bar{\Delta}, \bar{\Delta}'$ we take each interval of Δ, in (a', b'), as an interval of $\bar{\Delta}$, and each interval of Δ', in (a', b'), as an interval of $\bar{\Delta}'$. Further, we take for the parts of $\overline{\Delta}$, $\overline{\Delta}'$ in (a, a') and (b', b), the set of those intervals which are common to the parts of Δ and Δ' that lie in (a, a') and (b', b). In case a' is contained in intervals (α, β), (α', β') of Δ, and of Δ', respectively, we take (a', β), (a', β') as intervals of $\overline{\Delta}$ and $\overline{\Delta}'$ respectively, where $a' > a$. A similar specification will refer to b'.

---

* *Trans. Amer. Math Soc.* vol. II (1901), p. 296.

It is now clear that $f_{\Delta}(x) = f_{\Delta}(x)$, and $f_{\Delta'}(x) = f_{\overline{\Delta}'}(x)$, when x is in (a', b'); and that, in (a, a') or in (b', b), we have $f_{\overline{\Delta}}(x) = f_{\overline{\Delta}'}(x)$. It follows that

$$\int_a^b f_{\Delta}(x)\,dx - \int_a^b f_{\overline{\Delta}'}(x)\,dx = \int_{a'}^{b'} f_{\overline{\Delta}}(x)\,dx - \int_{a'}^{b'} f_{\Delta'}(x)\,dx,$$

and hence that
$$\left|\int_a^b f_{\Delta}(x)\,dx - \int_a^b f_{\overline{\Delta}'}(x)\,dx\right| \geqq \epsilon.$$

Moreover it is clear, from the mode of construction of $\overline{\Delta}$, $\overline{\Delta}'$, that

$$m(\overline{\Delta}) < \zeta_\epsilon \text{ and } m(\overline{\Delta}') < \zeta_\epsilon.$$

The impossibility in question has therefore been demonstrated.

Since, for every pair of numbers a', b', such that $a \leq a' < b' \leq b$, corresponding to an arbitrarily chosen ϵ, the number $\frac{1}{2}\zeta_\epsilon$ can be so chosen that

$$\left|\int_{a'}^{b'} f_{\Delta}(x)\,dx - \int_{a'}^{b'} f_{\Delta'}(x)\,dx\right| < \epsilon$$

for every pair of sets Δ, Δ' that enclose the set H narrowly, and such that $m(\Delta) < \frac{1}{2}\zeta_\epsilon$, $m(\Delta') < \frac{1}{2}\zeta_\epsilon$, it follows that $\int_{a'}^{b'} f(x)\,dx$ exists.

Moreover, since ζ_ϵ is independent of a' and b', we have established the following theorem.

If $\int_a^b f(x)\,dx$ *exists as an HL-integral, then* $\int_{a'}^{b'} f(x)\,dx$ *also exists, where* $a \leq a' < b' \leq b$; *and the convergence of this integral is uniform with respect to* a' *and* b'.

The last part of the theorem expresses the fact that

$$\left|\int_{a'}^{b'} f(x)\,dx - \int_{a'}^{b'} f_{\Delta}(x)\,dx\right| < \epsilon,$$

provided $m(\Delta) < \zeta_\epsilon$, for every value of a' and b', the number ζ_ϵ depending on ϵ.

454. It will be shewn that:

For the HL-integral, the theorem

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

is valid.

This follows from the corresponding theorem for the L-integral of $f_{\Delta}(x)$. For it appears that, employing the last theorem, the expressions on the two sides of the equation differ from one another by less than 2ϵ: and since ϵ is arbitrary, their equality is established.

Since the existence of the *HL*-integral of $f(x)$ in any sub-interval (a', b'), of (a, b), has been shewn to be a necessary consequence of the existence of the *HL*-integral over (a, b), it is clear that the integral of $f(x)$ taken over any finite set of non-overlapping intervals, contained in (a, b), also exists; being the sum of the integrals taken over the separate intervals. However, if a non-finite set of non-overlapping intervals be taken in (a, b), it is not in general true that the sum of the integrals taken over these intervals converges to a definite number, unless the integral of $f(x)$ is absolutely convergent, which case has been treated in connection with the *L*-integrals. It will in fact be shewn, by means of an example, that the property in question, that $f(x)$ is integrable over every non-finite set of intervals in (a, b), does not appertain to non-absolutely convergent integrals, and must be regarded as peculiar to absolutely convergent integrals. This does not however seem to be a sufficient reason for refraining from applying the term "integral" to non-absolutely convergent integrals, as defined above.

455. The following property of an indefinite *HL*-integral will be established:

The HL-integral $\int_a^x f(x)\,dx$ *is a continuous function of the upper limit* x.

Denoting the indefinite integral by $F(x)$, we have, in accordance with the addition theorem proved in § 454,

$$F(x+h) - F(x) = \int_x^{x+h} f(x)\,dx.$$

Employing the function $f_\Delta(x)$, which vanishes at all points of the intervals of the set Δ enclosing the set of points H, we have

$$\left|\int_x^{x+h} f(x)\,dx - \int_x^{x+h} f_\Delta(x)\,dx\right| < \epsilon,$$

provided $m(\Delta) < \zeta_\epsilon$, for every point x, and for all values of h, such that $x + h$ is in (a, b). The integral $\int_a^x f_\Delta(x)\,dx$ being a continuous function of x, the numerical value of $\int_x^{x+h} f_\Delta(x)\,dx$ in the interval $(0, h_1)$, of h, is less than ϵ, if h_1 be properly chosen. Hence the numerical value of $\int_x^{x+h} f(x)\,dx$ is less than 2ϵ, if h is in the interval $(0, h_1)$. Since ϵ is arbitrary, this shews that $F(x)$ is a continuous function of x, on the right, in the interval (a, b). Since $F(x) - F(x-h) = \int_{x-h}^{x} f(x)\,dx$, it can be shewn, in a similar manner, that $F(x)$ is continuous at x, on the left.

In an interval of the set $C(\Delta)$, $\int_a^x f(x)\,dx$ differs by a constant from an L-integral, and therefore $\int_a^c f(x)\,dx$ possesses, almost everywhere in the interval, a differential coefficient equal to $f(x)$. Since the measure of $C(\Delta)$ converges to $b-a$, when Δ belongs to a sequence $\{\Delta_n\}$, such that $m(\Delta_n)$ converges to zero, we have the following theorem, which is the extension to the HL-integral, of the property of the L-integral proved in § 405.

The indefinite HL-integral $\int_a^x f(x)\,dx$ *has, almost everywhere in* (a, b), *a differential coefficient of which the value is* $f(x)$.

THE HL-INTEGRAL OVER A FINITE SET OF INTERVALS

456. An extension of the theorem that the integral of $f(x)$ over any interval contained in (a, b) necessarily exists, if the HL-integral over (a, b) exists, will now be made.

Let D denote a finite set of intervals, each one of which is contained in an interval complementary to the set H, and no two of which are contained in any one such interval.

Using the notation of § 453, as regards $\Delta^{(1)}, \Delta^{(2)}$, such that $m(\Delta^{(1)}) < \frac{1}{2}\delta_\epsilon$, $m(\Delta^{(2)}) < \frac{1}{2}\delta_\epsilon$, it will be shewn that

$$\left|\int_{(D)} f_{\Delta^{(1)}}(x)\,dx - \int_{(D)} f_{\Delta^{(2)}}(x)\,dx\right| < \epsilon.$$

For, assuming that the expression on the left-hand side is $\geq \epsilon$, we can define $\Delta^{(3)}, \Delta^{(4)}$ as follows:

The parts of $\Delta^{(3)}, \Delta^{(4)}$ that are interior to $C(D)$, which is a finite set of intervals, we take to consist, in each case, of those intervals that are common to parts of $\Delta^{(1)}$ and $\Delta^{(2)}$ interior to $C(D)$. If intervals (α, β) (α', β'), of $\Delta^{(1)}$ and $\Delta^{(2)}$, contain an end-point a_ν, of an interval of D, we take (α', β), (α', β'), when $\alpha' > \alpha$, or (α, β), (α', β), if $\alpha' < \alpha$, as intervals of $\Delta^{(3)}$ and $\Delta^{(4)}$ respectively. No interval of $\Delta^{(1)}$ or $\Delta^{(2)}$ is interior to an interval of D.

We have now $m(\Delta^{(3)}) < \frac{1}{2}\delta_\epsilon$, $m(\Delta^{(4)}) < \frac{1}{2}\delta_\epsilon$; and

$$f_{\Delta^{(3)}}(x) = f_{\Delta^{(1)}}(x), \quad f_{\Delta^{(4)}}(x) = f_{\Delta^{(2)}}(x),$$

where x is in D; also, in $C(D)$, we have $f_{\Delta^{(3)}}(x) = f_{\Delta^{(4)}}(x)$; it then follows that

$$\left|\int_a^b f_{\Delta^{(3)}}(x)\,dx - \int_a^b f_{\Delta^{(4)}}(x)\,dx\right| \geq \epsilon,$$

which is inconsistent with the conditions that

$$m(\Delta^{(3)}) < \tfrac{1}{2}\delta_\epsilon, \quad m(\Delta^{(4)}) < \tfrac{1}{2}\delta_\epsilon.$$

Hence
$$\left|\int_{(D)} f_{\Delta^{(1)}}(x)\,dx - \int_{(D)} f_{\Delta^{(2)}}(x)\,dx\right| < \epsilon.$$

It follows that $\left|\int_{(D)} f(x)\,dx - \int_{(D)} f_\Delta(x)\,dx\right| < \epsilon$, provided $m(\Delta) < \frac{1}{2}\delta_\epsilon$. It has thus been established that:

If D denote a finite set of intervals, each of which is in one of the complementary intervals of H, and no two of which are in one and the same such complementary interval, then if ϵ is prescribed, there exists a number $\frac{1}{2}\delta_\epsilon$, such that

$$\left|\int_{(D)} f(x)\,dx - \int_{(D)} f_\Delta(x)\,dx\right| < \epsilon$$

for every such set D, and every set Δ enclosing H narrowly, such that

$$m(\Delta) < \tfrac{1}{2}\delta_\epsilon.$$

This is a particular case of a somewhat more general theorem given by E. H. Moore, which allows more latitude as regards the set of intervals D.

Since the function $f_\Delta(x)$ is summable, we see that, in accordance with the property of summable functions established in § 392,

$$\left|\int_{(D)} f_\Delta(x)\,dx\right| < \epsilon,$$

provided $m(D)$ is less than some number δ_ϵ' which converges to zero with ϵ.

Also, provided D satisfies the condition of the above theorem, we have

$$\left|\int_{(D)} f(x)\,dx - \int_{(D)} f_\Delta(x)\,dx\right| < \epsilon.$$

It now follows that $\left|\int_{(D)} f(x)\,dx\right| < 2\epsilon$.

We have now the theorem that:

If D be a set of intervals, finite in number, each of which is in an interval complementary to the set H, of points of non-summability, and no two of which are in the same such interval, then

$$\left|\int_{(D)} f(x)\,dx\right| < \epsilon$$

for all such sets D, provided $m(D) < \delta_\epsilon''$; where ϵ is arbitrary, and δ_ϵ'' depends upon ϵ, and converges to zero as ϵ does so.

THE CONDITIONS FOR THE EXISTENCE OF AN *HL*-INTEGRAL

457. The following theorem, due to E. H. Moore*, contains the necessary and sufficient conditions for the existence of the *HL*-integral of a function $f(x)$, defined in the linear interval (a, b), in which H is the set of points of non-summability of $f(x)$. The content of H is assumed to be zero:

The complementary intervals of H being denoted by (a_ν, b_ν), the necessary and sufficient conditions for the existence of $\int_a^b f(x)\,dx$ are:

---

* *Trans. Amer. Math. Soc.* vol. II (1901), p. 324.

(1), *that all the integrals* $\int_{a_\nu}^{b_\nu} f(x)\,dx$ *shall exist, each such integral being defined as the limit of* $\int_{a_\nu+\epsilon}^{b_\nu-\epsilon'} f(x)\,dx$, *when* ϵ, ϵ' *converge independently to the limit zero, and,*

(2), *that* $\omega_1 + \omega_2 + \dots + \omega_\nu$ *shall converge to a definite number, as* ν *is indefinitely increased; where* ω_ν *denotes the fluctuation of* $\int_{a_\nu}^{x} f(x)\,dx$ *in the interval* (a_ν, b_ν).

Moreover, when the conditions (1) *and* (2) *are satisfied, the sum*

$$\sum_{r=1}^{\nu} \int_{a_r}^{b_r} f(x)\,dx$$

is convergent, and its limit, as $\nu \sim \infty$, *is* $\int_a^b f(x)\,dx$.

To shew that the conditions are necessary, we assume that $\int_a^b f(x)\,dx$ exists; it then follows from the theorem of § 453, that $\int_{a_\nu}^{b_\nu} f(x)\,dx$ exists. If ξ_ν', ξ_ν'' be the points of (a_ν, b_ν) at which $\int_{a_\nu}^{x} f(x)\,dx$ attains its maximum and minimum values we have $\omega_\nu = \int_{\xi_\nu''}^{\xi_\nu'} f(x)\,dx$.

The number μ may be chosen so large that $\sum_{\nu=\mu+1}^{\infty} (b_\nu - a_\nu) < \delta_\epsilon''$. For some values of ν we may have $\xi_\nu' > \xi_\nu''$, and for others $\xi_\nu' < \xi_\nu''$; one, or both, of these two sets of values of ν will be an infinite set. If $\Sigma\omega_\nu$ is not convergent, a number μ' ($> \mu + 1$) can be found such that $\sum_{\nu=\mu+1}^{\mu'} \omega_\nu > 2\epsilon$. One at least of the two sets of those values of ν between $\mu + 1$ and μ' for which $\xi_\nu' > \xi_\nu''$, and for which $\xi_\nu' < \xi_\nu''$, must be such that the sum of the corresponding values of ω_ν is $> \epsilon$, and the sum of the corresponding intervals $| \xi_\nu'' - \xi_\nu' |$ is $< \delta_\epsilon''$. We have thus a finite set of intervals D, whose sum is $< \delta_\epsilon''$, all contained in complementary intervals of G, no two of which are in the same such interval, and such that

$$\left| \int_{(D)} f(x)\,dx \right| > \epsilon,$$

which is contrary to the last theorem of § 456.

It follows that $\Sigma\omega_\nu$ must converge.

Further, let (α_ν, β_ν) be interior to (a_ν, b_ν) and such that

$$\beta_\nu - \alpha_\nu = \left(1 - \frac{1}{k}\right)(b_\nu - a_\nu).$$

We may choose m so that $\sum_{\nu=1}^{\nu=m}(b_\nu - a_\nu) > b - a - \eta$, where η is positive, and arbitrarily chosen. Then

$$\sum_{\nu=1}^{\nu=m}(\beta_\nu - \alpha_\nu) > (b - a - \eta)\left(1 - \frac{1}{k}\right) > b - a - \eta - \frac{1}{k}(b - a).$$

If we take Δ to be the set of intervals complementary to (α_ν, β_ν), for $\nu = 1, 2, 3, \ldots m$, we have $m(\Delta) < \eta + \frac{1}{k}(b - a) < \delta$, if η and k are properly chosen. It then follows that

$$\left|\int_a^b f(x)\,dx - \sum_{\nu=1}^{\nu=m}\int_{\alpha_\nu}^{\beta_\nu} f(x)\,dx\right| < \epsilon.$$

Also k may be so chosen that $\sum_{\nu=1}^{\nu=m}\int_{\alpha_\nu}^{\beta_\nu} f(x)\,dx$ differs from $\sum_{\nu=1}^{\nu=m}\int_{a_\nu}^{b_\nu} f(x)\,dx$ by less than ϵ; we have then

$$\left|\int_a^b f(x)\,dx - \sum_{\nu=1}^{\nu=m}\int_{a_\nu}^{b_\nu} f(x)\,dx\right| < 2\epsilon.$$

It now follows that $\sum_{\nu=1}^{\infty}\int_{a_\nu}^{b_\nu} f(x)\,dx$ converges to the value of $\int_a^b f(x)\,dx$.

To prove that the conditions (1) and (2) are sufficient; let ν' be such that $\sum_{\nu'+1}^{\infty}\omega_\nu < \frac{1}{2}\epsilon$; then if (α_ν, β_ν) denote any interval contained in (a_ν, b_ν), we have

$$\left|\sum_{\nu=\nu'+1}^{\infty}\int_{\alpha_\nu}^{\beta_\nu} f(x)\,dx\right| < \tfrac{1}{2}\epsilon.$$

Let the finite set Δ enclose H narrowly; the complementary set $C(\Delta)$ is a finite set of intervals interior to a finite set of the intervals $C(H)$ complementary to H. Assume that these intervals $C(H)$ are arranged in descending order of magnitude. Those of them that are of length $> m(\Delta)$ must each contain an interval of $C(\Delta)$, and others may also contain intervals of $C(\Delta)$. Let the first s of the intervals (a_ν, b_ν) be each of length $> m(\Delta)$. The convergence of $\sum_{\nu=1}^{\infty}\int_{a_\nu}^{b_\nu} f(x)\,dx$ follows from that of $\Sigma\omega_\nu$. We have then

$$\left|\int_a^b f_\Delta(x)\,dx - \sum_{\nu=1}^{\infty}\int_{a_\nu}^{b_\nu} f(x)\,dx\right|$$

$$\leq \sum_{\nu=1}^{\nu=s}\left|\int_{a_\nu}^{b_\nu} f(x)\,dx - \int_{\alpha_\nu}^{\beta_\nu} f(x)\,dx\right| + \sum_{\nu=s+1}^{\infty}\left|\int_{a_\nu}^{b_\nu} f(x)\,dx\right| + \sum_{n'>s}\left|\int_{\alpha_{n'}}^{\beta_{n'}} f(x)\,dx\right|,$$

where n' $(> s)$ has a finite set of values, and (α_ν, β_ν) are the intervals of $C(\Delta)$.

We may suppose s to be so chosen that $\sum_{\nu=s}^{\infty}\omega_\nu < \epsilon$, and thus that

$$\sum_{n'>s}\left|\int_{\alpha_{n'}}^{\beta_{n'}} f(x)\,dx\right| < \epsilon,$$

and also so that $\sum_{\nu=s+1}^{\infty}\left|\int_{a_\nu}^{b_\nu} f(x)\,dx\right| < \epsilon$. The number s having been so chosen, we may suppose Δ so chosen that $m(\Delta)$ is less than each of the first s intervals of $C(H)$. Also $m(\Delta)$ may be chosen, still smaller if necessary, so that

$$\sum_{\nu=1}^{\nu=s}\left|\int_{a_\nu}^{b_\nu} f(x)\,dx - \int_{\alpha_\nu}^{\beta_\nu} f(x)\,dx\right| < \epsilon.$$

We now have $\left|\int_a^b f_\Delta(x)\,dx - \sum_{\nu=1}^{\infty}\int_{a_\nu}^{b_\nu} f(x)\,dx\right| < 3\epsilon$, provided $m(\Delta)$ is less than some number depending on ϵ; and since this holds for every value of ϵ, we see that $\int_a^b f_\Delta(x)\,dx$ must converge to the value of

$$\sum_{\nu=1}^{\infty}\int_{a_\nu}^{b_\nu} f(x)\,dx,$$

as $m(\Delta) \sim 0$; and thus the existence of $\int_a^b f(x)\,dx$ has been established.

It has been here established that, if $\int_a^b f(x)\,dx$ exists as an HL-integral, it is necessary that $\Sigma\int_{a_\nu}^{b_\nu} f(x)\,dx$ should converge, where (a_ν, b_ν) denotes a contiguous interval of the set H of points of non-summability of $f(x)$, it being assumed that $m(H) = 0$, and $\int_{a_\nu}^{b_\nu} f(x)\,dx$ being determined as $\lim\limits_{\substack{\alpha\sim a_\nu\\ \beta\sim b_\nu}}\int_\alpha^\beta f(x)\,dx$, where (α, β) is interior to (a_ν, b_ν), but it is also necessary to assume that, ω_ν denoting the fluctuation of $\int_{a_\nu}^{x} f(x)\,dx$ in (a_ν, b_ν), the series of positive terms $\Sigma\omega_\nu$ should be convergent.

458. A more general definition of $\int_a^b f(x)\,dx$ has been suggested by Lebesgue*, as a development of a definition given by Jordan. This is obtained by substituting for the condition that $\Sigma\omega_n$ should converge, the less stringent condition that $\Sigma\left|\int_{a_n}^{b_n} f(x)\,dx\right|$ should be convergent. It is clear that an HL-integral is also an integral in accordance with this definition of Lebesgue, but the converse is not necessarily the case. It can be seen that the substitution of Lebesgue's definition for that of Harnack involves the introduction of a limitation in the selection of the sets of intervals Δ that include the points of H.

---

* See his "Remarques sur les théories de la mesure et de l'intégration," *Annales sc. de l'école normale* (3), vol. xxxv (1918), p. 204.

This limitation must be such that, in the proof of sufficiency in E. H. Moore's theorem, the convergence of $\sum_{\nu=1} \left| \int_{\alpha_\nu}^{\beta_\nu} f(x)\, dx \right|$ can be inferred from that of $\sum_{\nu=1} \left| \int_{a_\nu}^{b_\nu} f(x)\, dx \right|$.

In case the sum of the integrals over the intervals (a_ν, b_ν) were not absolutely convergent, the sum of the integrals themselves, if convergent, would have a sum which would depend upon the order in which the intervals are taken, and thus would not afford a suitable definition of the integral over (a, b).

459. A method will now be given of constructing a function $f(x)$ which is continuous at every point of the interval (a, b), except at the point b, at which the function has an infinite discontinuity of such a character that $\int_a^b f(x)\, dx$ converges non-absolutely.

Let a sequence of intervals $(a_1, b_1), (a_2, b_2), \dots (a_n, b_n), \dots$ be defined in the interval (a, b), such that no two of them overlap, and that b is the limiting point of each of the sequences

$$(a_1, a_2, \dots a_n, \dots), \quad (b_1, b_2, \dots b_n, \dots).$$

Let $u_1 + u_2 + \dots + u_n + \dots$ denote a non-absolutely convergent arithmetic series; that is, the series is to be convergent, but not the series

$$| u_1 | + | u_2 | + \dots + | u_n | + \dots.$$

In (a_n, b_n), let $f(x)$ be defined so as to be continuous, and everywhere of the same sign, and let $f(x)$ be zero at a_n and b_n. Further, let $f(x)$ be so chosen, in (a_n, b_n), that $\int_{a_n}^{b_n} f(x)\, dx = u_n$. At all points of (a, b), external to all the intervals (a_n, b_n), let $f(x) = 0$.

The function $f(x)$, so defined in (a, b), is continuous, except at the point b.

In (a_n, b_n), the function $| f(x) |$ has a maximum value greater than $| u_n |/(b_n - a_n)$, and therefore $f(x)$ has indefinitely great positive and negative values in every neighbourhood of the point b.

For, if there existed a positive number k, such that $| u_n |/(b_n - a_n) < k$, for all values of n, we should have $\sum_{r=1}^{r=n} | u_n | < k \sum_{r=1}^{r=n} (b_r - a_r)$, and thus the series $\sum_{r=1}^{\infty} | u_n |$ would be convergent, which is not the case.

We have now
$$\int_a^x f(x)\, dx = \sum_{r=1}^{r=n} u_n + \theta u_{n+1},$$
if x lies in the interval (b_n, b_{n+1}), where $0 \leq \theta \leq 1$.

Also the integral $\int_a^b f(x)\,dx$ is defined as $\lim\limits_{x\sim b}\int_a^x f(x)\,dx$; and its value is therefore $\lim\limits_{n\sim\infty}\sum\limits_{r=1}^{r=n} u_n$, which, by hypothesis, has a definite value.

It is further clear that $\int_a^b |f(x)|\,dx$ is not finite, since the series $|u_1| + |u_2| + \dots + |u_n| + \dots$ is not convergent.

This result may be employed to illustrate the fact that the non-absolutely convergent integral is not necessarily the limit of the sum of the integrals taken over a set of intervals which, in the limit, converges to the whole interval of integration; and thus that such an integral is not a broad integral.

Let the integral of $f(x)$ be taken over the intervals (a, b_m), (a_{p_1}, b_{p_1}), (a_{p_2}, b_{p_2}), … (a_{p_r}, b_{p_r}), where $p_1, p_2, \dots p_r$ are increasing integers, all greater than m, and such that $u_{p_1}, u_{p_2}, \dots u_{p_r}$ are all of the same sign. It is clear that m may be so chosen that $\int_a^{b_m} f(x)\,dx$ is arbitrarily near in value to $\int_a^b f(x)\,dx$; then, for such a value of m, the integers $p_1, p_2, \dots p_r$ may be so chosen that $u_{p_1} + u_{p_2} + \dots + u_{p_r}$ is as large as we please, since the series Σu_n does not converge absolutely. As m is increased indefinitely, the set of intervals (a, b_m), (a_{p_1}, b_{p_1}), … (a_{p_r}, b_{p_r}) converges to the whole interval (a, b), the complementary part of (a, b) diminishing indefinitely, and yet the sum of the integrals of $f(x)$ taken over the intervals of the set does not converge.

460. The construction here given of an *HL*-integral, with a single point of non-summability, may be employed to illustrate the fact that, in the definition of such an integral given in § 453, the condition is indispensable that the set Δ of intervals must be such that there is at least one point of non-summability in each interval of Δ. This fact differentiates the non-absolutely convergent *HL*-integral from an absolutely convergent integral, in which $\int_{C(\Delta)} f(x)\,dx$ converges to $\int_a^b f(x)\,dx$, for every set of intervals Δ, when $m(\Delta) \sim 0$. It is not in fact true that, in defining the *HL*-integral, the set of points $a_1, a_2, \dots a_n, \dots b_1, b_2, \dots b_n, \dots b$, which is of content zero, may be excluded by enclosing these points in a set of intervals of arbitrarily small sum.

For we may include all the $2m$ points $a_1, a_2, \dots a_m, b_1, b_2, \dots b_m$ which occur in (a, b_m) in a finite set of intervals, so that when these are excluded from the interval (a, b_m), the integral $\int_a^{b_m} f(x)\,dx$ is altered by an arbitrarily small amount. Again, we may shorten each of the intervals (a_{p_1}, b_{p_1}), (a_{p_2}, b_{p_2}), … (a_{p_r}, b_{p_r}) at each end, so that the sum of the integrals of $f(x)$

taken over these intervals is diminished by an arbitrarily small amount. All the points $a_1, a_2, \dots b_1, b_2, \dots$ are now included in intervals of a finite set, such that the integral of $f(x)$ over the complementary intervals is arbitrarily great. These complementary intervals consist of those intervals which have been obtained by shortening the intervals (a_{p_1}, b_{p_1}), (a_{p_2}, b_{p_2}), … (a_{p_r}, b_{p_r}), and of the parts of (a, b_m) which remain when the points

$$a_1, b_1, a_2, b_2, \dots a_m, b_m$$

have been included in a suitable set of intervals.

Let $\phi(x)$ be a function for which $\int_a^b \phi(x)\,dx$ exists as an HL-integral, with the points $a_1, a_2, \dots a_n, \dots b_1, b_2, \dots b_n, \dots b$, as the set of points of infinite discontinuity, in the neighbourhoods of which $\phi(x)$ is not summable. Also the integral $\int_a^b f(x)\,dx$ exists, as constructed above, with its single point of non-summability at b. It appears however that

$$\int_a^b \{f(x) + \phi(x)\}\,dx$$

does not exist as an HL-integral. The set H, of points of non-summability of the function $f(x) + \phi(x)$, consists of the points

$$a_1, b_1, a_2, b_2, \dots a_n, b_n, \dots,$$

and in general of the point b.

We may now define a set Δ, of intervals enclosing the points of H, to consist of the finite set defined above for the case of the integral of $f(x)$. For a sequence of such sets Δ, $\int_{C(\Delta)} f(x)\,dx$ increases indefinitely, as $m(\Delta) \sim 0$, whereas $\int_{C(\Delta)} \phi(x)\,dx$ has a finite limit; thus

$$\int_{C(\Delta)} \{f(x) + \phi(x)\}\,dx$$

increases indefinitely as $m(\Delta) \sim 0$, and therefore $\int_a^b \{f(x) + \phi(x)\}\,dx$ does not exist. It will however appear in §§ 472, 473 that the integral exists as a D-integral.

It will be proved that.

If $f(x)$, $\phi(x)$ have HL-integrals in (a, b), and the sets H_1, H_2, of points of non-summability of the two functions have no point in common, then $f(x) + \phi(x)$ has an HL-integral in (a, b), given by

$$\int_a^b \{f(x) + \phi(x)\}\,dx = \int_a^b f(x)\,dx + \int_a^b \phi(x)\,dx.$$

Denoting $f(x) + \phi(x)$ by $\psi(x)$, it is clear that the set of points of non-summability of $\psi(x)$ is $K \equiv H_1 + H_2$. The two closed sets H_1, H_2 have a finite distance between them, and therefore H_2 must be contained within the intervals D_2 of a finite set, each interval of which is within one of the intervals contiguous to H_1. The part of H_2 in each such interval is closed. If the set K is enclosed narrowly in a finite set of intervals Δ, then, provided $m(\Delta)$ is sufficiently small, Δ is the sum of a set Δ_1 of intervals enclosing narrowly the points of H_1, and of a set Δ_2 of intervals contained in D_2, enclosing narrowly the points of H_2. The integrals $\int_a^b f_\Delta(x)\,dx$, $\int_a^b f_{\Delta_1}(x)\,dx$ differ from one another by $\int_{(\Delta_2)} f(x)\,dx$, and this is an L-integral which converges to zero with $m(\Delta)$; therefore, in defining $\int_a^b f(x)\,dx$, we may employ the set K instead of the set H_1. Similarly, it is seen that, in defining $\int_a^b \phi(x)\,dx$, we may employ the set K instead of H_2.

Since
$$\int_a^b \psi_\Delta(x)\,dx = \int_{C(\Delta)} f(x)\,dx + \int_{C(\Delta)} \phi(x)\,dx$$
and since the limits of the integrals on the right-hand side, as $m(\Delta) \sim 0$, are the HL-integrals of $f(x)$ and $\phi(x)$ respectively, the truth of the theorem follows.

When H_1 and H_2 have points in common, such a point may or may not be a point of non-summability of $f(x) + \phi(x)$. It can be shewn that:

If $f(x)$, $\phi(x)$ both have HL-integrals in (a, b), and have one and the same set H of points of non-summability, then if the set of points of non-summability of $f(x) + \phi(x)$ consists of the points of H with the exception of those in a closed set L contained in H (which set L may be finite or absent), then $f(x) + \phi(x)$ has an HL-integral in (a, b), given by the sum of the HL-integrals of $f(x)$ and $\phi(x)$.

The sets $H - L$ and L both being closed, it follows, as above, that L is contained within intervals of a finite set D, each interval of which is interior to one of the intervals contiguous to $H - L$. It can then be shewn that, in defining the integral of $\psi(x)$, the set H may be employed instead of the set $H - L$. For $\int_a^b \psi_\Delta(x)\,dx$, $\int_a^b \psi_{\Delta_1}(x)\,dx$ differ from one another by the L-integral of $\psi(x)$ taken over the set Δ_2; where $\Delta = \Delta_1 + \Delta_2$, and Δ_1, Δ_2 enclose narrowly the points of the sets $H - L$ and L respectively. The result then follows as in the last case.

461. If $\int_a^b \phi(x)\,dx$ is an HL-integral, for the function $\phi(x)$, and the set of points of non-summability of $\phi(x)$ is H, and $f(x)$ is bounded and summable in (a, b), it may happen that some or all of the points of H are not points of non-summability of the product function $f(x)\,\phi(x)$. This will be the case, for example, if $f(x)$ has the value zero in a neighbourhood of a point of H. If, at any point ξ, of H, $f(\xi + 0)$ exists and is numerically greater than some positive number α, we have, provided ϵ is sufficiently small,

$$\int_{\xi}^{\xi+\epsilon} | f(x)\, \phi(x) |\, dx > \alpha \int_{\xi}^{\xi+\epsilon} | \phi(x) |\, dx.$$

It follows that if ξ is a point of non-summability on the right, of $\phi(x)$, then it is also a point of non-summability of $f(x)\, \phi(x)$. A similar remark applies to the left of the point ξ.

The following property* of an HL-integral will now be established

If $\int_a^b \phi(x)\, dx$ is an HL-integral, the set of points of non-summability of $\phi(x)$ being H, and if $f(x)$ be a function with bounded variation in (a, b), such that $f(x)\, \phi(x)$ is non-summable at every point of H, with the possible exception of those belonging to a closed set K, of points of H, then

$$\int_a^b f(x)\, \phi(x)\, dx$$

exists as an HL-integral. The set K may be finite, or absent.

The condition in the theorem is certainly satisfied, in particular, if there are only a finite number of points of H at which $f(x)$ is continuous and has the value zero. An important example of the theorem is that, if $\phi(x)$ has an HL-integral in (a, b), then $\phi(x) \cos \lambda x$, $\phi(x) \sin \lambda x$ have HL-integrals in the interval, λ denoting any constant. For the only points of H which may not be points of infinite discontinuity of the product functions are the zeros of $\cos \lambda x$, or of $\sin \lambda x$.

First, let it be assumed that all the points of H are points of non-summability of $f(x)\, \phi(x)$. Let Δ, Δ' be two sets of intervals enclosing narrowly the points of H, and such that

$$\left| \int_a^x \phi_\Delta(x)\, dx - \int_a^x \phi_{\Delta'}(x)\, dx \right| < \epsilon,$$

where $m(\Delta) < \zeta_\epsilon$, $m(\Delta') < \zeta_\epsilon$, for all points x in (a, b).

Let $F(x) \equiv \phi_\Delta(x) - \phi_{\Delta'}(x)$, and suppose that $f(x) = f_1(x) - f_2(x)$, where $f_1(x)$, $f_2(x)$ are monotone bounded functions.

We have

$$\int_a^b F(x)\, f_1(x)\, dx = f_1(a) \int_a^{\xi} F(x)\, dx + f_1(b) \int_{\xi}^{b} F(x)\, dx,$$

with a similar equation for $f_2(x)$; ξ is in each case some point in (a, b). From this we have

$$\left| \int_a^b \phi_\Delta(x)\, f(x)\, dx - \int_a^b \phi_{\Delta'}(x)\, f(x)\, dx \right|$$
$$< 2\epsilon \{ | f_1(a) | + | f_1(b) | + | f_2(a) | + | f_2(b) | \}.$$

---

* See Hobson, *Proc. Lond. Math. Soc.* (2), vol VII (1909), p. 22.

Denoting the number on the right-hand side by ϵ', we see that, since ζ_ϵ converges to zero with ϵ', the condition for the existence of

$$\int_a^b f(x)\,\phi(x)\,dx$$

as an HL-integral is satisfied.

In case there exists a closed set K, of points of H, which are not points of non-summability of $f(x)\,\phi(x)$, we see, exactly as at the end of § 460, that in defining the HL-integral of $f(x)\,\phi(x)$ we may employ the set H, instead of the closed set $H - K$, of points of non-summability of the function. Thus the theorem is at once extended to this case. If H and K are identical, the HL-integral becomes an L-integral.

THE SECOND MEAN VALUE THEOREM FOR AN HL-INTEGRAL

462. With a view to the extension of the second mean value theorem given in § 422 to the case of HL-integrals, it will be first proved that:

If $\phi(x)$ has an HL-integral, and $f(x)$ is bounded and monotone in the interval (a, b), then $f(x)\,\phi(x)$ has an HL-integral, or an L-integral, in the same interval.

Let H and $\bar{H}$ be the sets of points of non-summability of the functions $\phi(x)$, $f(x)\,\phi(x)$ respectively. Any point x, of H, at which neither $f(x + 0)$ nor $f(x - 0)$ has the value zero is also a point of $\bar{H}$. If there be no line of invariability of $f(x)$ for which the function $f(x)$ has the value zero, there may be a single point c, in (a, b), at which $f(c + 0)$ or $f(c - 0)$ has the value zero, or at which both of them are zero. If c be a point of H it is not necessarily a point of $\bar{H}$, and in that case it is an isolated point of H. In accordance with the theorem of § 461, $f(x)\,\phi(x)$ has an HL-integral in (a, b), the set K consisting of the single point c. In defining the integral of $f(x)\,\phi(x)$, the set H may be employed instead of $\bar{H}$. In case there is a line of invariability (α, β) contained in (a, b), within which $f(x) = 0$, any points of H within (α, β) do not belong to $\bar{H}$. The end-point α, or the end-point β, in case it is a point of H, may, or may not, belong to $\bar{H}$. In the interval (α, β), $f(x)\,\phi(x)$ has an L-integral, and in the interval (a, α), or (β, b) it has an HL-integral, or an L-integral, according as the interval does or does not contain points of H. In any case it follows that $f(x)\,\phi(x)$ has an HL-integral (or an L-integral if all the points of H are within (α, β)) in the interval (a, b); and in defining the integral the set H may be employed.

It will now be proved that:

The second mean value theorem

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^X \phi(x)\,dx + B\int_X^b \phi(x)\,dx,$$

as given in § 422, *holds good when* $\phi(x)$ *has an* HL-*integral in* (a, b), *the function* $f(x)$ *being bounded and monotone in the same interval.*

In accordance with the last theorem, $\int_a^b f(x)\,\phi(x)\,dx$ exists as an HL-integral (or as an L-integral), and in its definition the set H, of points of non-summability of $\phi(x)$, may be employed, whether or not all the points of H are points of non-summability of $f(x)\,\phi(x)$.

We have

$$\int_a^b f(x)\,\phi_\Delta(x)\,dx = A\int_a^{X_\Delta}\phi_\Delta(x)\,dx + B\int_{X_\Delta}^b \phi_\Delta(x)\,dx,$$

where Δ is a set of intervals enclosing narrowly the set H, and X_Δ is some point in (a, b), dependent on Δ, A, B.

Now $\int_a^{X_\Delta}\phi_\Delta(x)\,dx - \int_a^{X_\Delta}\phi(x)\,dx$, $\int_{X_\Delta}^b \phi_\Delta(x)\,dx - \int_{X_\Delta}^b \phi(x)\,dx$ are both numerically less than an arbitrarily chosen positive number ϵ, provided $m(\Delta)$ is sufficiently small; this follows from the uniform convergence of $\int_a^x \phi_\Delta(x)\,dx$ to $\int_a^x \phi(x)\,dx$. Also $\int_a^b f(x)\,\phi_\Delta(x)\,dx$ differs from $\int_a^b f(x)\,\phi(x)\,dx$ by less than ϵ, if $m(\Delta)$ is sufficiently small. Hence we have

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^{X_\Delta}\phi(x)\,dx + B\int_{X_\Delta}^b \phi(x)\,dx + \eta,$$

where $|\eta|$ is arbitrarily small. From the continuity of

$$\int_a^{X_\Delta}\phi(x),\quad \int_{X_\Delta}^b \phi(x)\,dx,$$

we see, by reasoning similar to that in § 423, that

$$\int_a^b f(x)\,\phi(x)\,dx = A\int_a^X \phi(x)\,dx + B\int_X^b \phi(x)\,dx.$$

Bonnet's form of the mean value theorem may be deduced as in § 423.

It should be observed that the only case in which the monotone function $f(x)$ is such that the condition that all the points of H, with the possible exception of the points of a closed set contained in H, are points of non-summability of $f(x)\,\phi(x)$ may not be satisfied is when $f(x)$ is zero in the interior of an interval (α, β) contained in (a, b). For example, if β is a limiting point of H on both sides, and $f(\beta) = 0$, $f(\beta + 0) = 1$, the point β is a point of non-summability of $f(x)\,\phi(x)$, but the points of H on the left of β are points of summability of $f(x)\,\phi(x)$. Thus, in this case, the set of points of H which are points of summability of $f(x)\,\phi(x)$ is not a closed set; and thus the condition in the theorem of § 461 is not satisfied. In such a case, by changing $f(x)$ into $f(x) + c$, we have

$$\int_a^b [f(x) + c]\,\phi(x)\,dx = (A + c)\int_a^X \phi(x)\,dx + (B + c)\int_X^b \phi(x)\,dx,$$

the integral on the left-hand side existing as an HL-integral. We could only infer the existence of $\int_a^b f(x)\,\phi(x)\,dx$, by subtracting $\int_a^b c\phi(x)\,dx$ from both sides, and employing the theorem of § 454.

INTEGRATION BY PARTS FOR THE HARNACK-LEBESGUE INTEGRAL

463. The formula for integrating by parts, given in § 420, may be extended* to apply to the case in which one of the integrals, $\int_a^x U(x)\,dx$, which may be denoted by $u(x)$, is an HL-integral; and the other $\int_a^x V(x)\,dx$, or $v(x)$, is an L-integral.

Let $u_n(x) = \int_a^x U_n(x)\,dx$, where $U_n(x) = U(x)$, at every point not belonging to a finite set of intervals Δ, enclosing the points H, of non-summability of $U(x)$; and $U_n(x) = 0$, in Δ.

Since u_n and $u_n v$ are L-integrals, we have, since $\frac{du_n}{dx}$, $\frac{dv}{dx}$ exist almost everywhere (see § 455),

$$[u_n v]_a^b = \int_a^b v\frac{du_n}{dx}\,dx + \int_a^b u_n\frac{dv}{dx}\,dx.$$

As $n \sim \infty$, so that $m(\Delta) \sim 0$, we have $[u_n v]_a^b \sim [uv]_a^b$.

Also $$\left|\int_a^b u\frac{dv}{dx}\,dx - \int_a^b u_n\frac{dv}{dx}\,dx\right| < \epsilon\int_a^b \left|\frac{dv}{dx}\right| dx,$$

because, for sufficiently large values of n, we have $|u(x) - u_n(x)| < \epsilon$, whatever value x may have, in accordance with the theorem of § 453. It follows that

$$\lim_{n\sim\infty}\int_a^b u_n\frac{dv}{dx}\,dx = \int_a^b u\frac{dv}{dx}\,dx,$$

and hence, utilizing the above equation, we see that $\lim\limits_{n\sim\infty}\int_a^b v\frac{du_n}{dx}\,dx$ exists and is equal to $[uv]_a^b - \int_a^b u\frac{dv}{dx}\,dx$.

Thus the equation

$$\int_a^b V(x)\left\{\int_a^x U(x)\,dx\right\}dx = \int_a^b U(x)\,dx\int_a^b V(x)\,dx - \int_a^b U(x)\left\{\int_a^x V(x)\,dx\right\}dx$$

holds good, when one of the two indefinite integrals

$$\int_a^x U(x)\,dx,\quad \int_a^x V(x)\,dx,$$

exists only as an HL-integral, the other being an L-integral.

---

* W. H. Young, *Proc. Lond. Math. Soc.* ser. 2, vol. IX, p. 432.

THE DENJOY INTEGRAL

464. In the definition of the HL-integral, even if it be modified in the manner proposed by Lebesgue (§ 458), it is assumed that the set H, of points of non-summability of the function, has content zero. A definition has been introduced* by Denjoy, which is applicable when the set H is not so restricted, but is capable of being any non-dense closed set. The process of passing from a given function $f(x)$, defined in the linear interval (a, b), to a function $F(x)$, which has a relation to $f(x)$ similar to that of the indefinite L-integral of a summable function to the function itself, has been termed by Denjoy *totalization* of $f(x)$; it will however be here spoken of as integration (D), and the function $F(x)$ will be spoken of as the indefinite D-integral of $f(x)$, whenever it exists. Thus, in any interval (α, β), of (a, b), the D-integral of $f(x)$ will be denoted by $\int_\alpha^\beta f(x)\,dx$, or by $F(\beta) - F(\alpha)$, or also by $V(\alpha, \beta)$.

The definition of the Denjoy Integral, or D-integral, in the most general case, can only be given by indicating the steps of a gradual process by which, commencing with the employment of L-integrals, a function is obtained which satisfies certain postulations which are of the nature of definitions. These may be stated as follows:

Having given a measurable function $f(x)$, defined for the interval (a, b),

(1). *In any interval* (α, β), *contained in* (a, b), *in which* $f(x)$ *is integrable* (L), $V(\alpha, \beta)$ *is taken to be the L-integral* $\int_\alpha^\beta f(x)\,dx$.

(2). *For a finite set of intervals* (α_1, α_2), (α_2, α_3), ... (α_{n-1}, α_n), *each one of which abuts on the next, and for which* $V(\alpha_1, \alpha_2)$, $V(\alpha_2, \alpha_3)$, ... $V(\alpha_{n-1}, \alpha_n)$ *have been defined,* $V(\alpha_1, \alpha_n)$ *is defined as* $\sum_{r=1}^{r=n-1} V(\alpha_r, \alpha_{r+1})$.

(3). *If* (α, β) *be any interval in* (a, b), *and* $V(\alpha', \beta')$ *has been defined for every interval* (α', β') *interior to* (α, β), *then* $V(\alpha, \beta)$ *is taken to be the limit of* $V(\alpha', \beta')$, *when* $\alpha' \sim \alpha$, $\beta' \sim \beta$, *independently of one another, it being assumed that this limit exists.*

(4). *Let P be a perfect set of points in an interval* (α, β) *contained in* (a, b), *and assume that* $f(x)$ *is summable over the set P, and suppose moreover that* $V(\alpha', \beta')$ *has been defined for every interval* (α', β'), *of* (α, β), *which contains no point of P as an interior point. Let* (α_n, β_n), *where* $n = 1, 2, 3, \ldots,$

---

* *Comptes Rendus*, Paris, vol. CLIV (1912), p. 859 and p. 1075. See also Lusin, *ibid.* vol. XLV, p. 474. Denjoy's detailed investigations are given in four memoirs referred to on p. 715. See also a thesis by Nalli, *Esposizione e confronto critico delle diverse definizioni di una funzione limitata o no*, Palermo, 1914 For an extension of Denjoy's method of integration to the case of the derivatives of functions of two or more variables, see Looman, *Fundamenta Math.* vol. IV (1923), p. 246.

denote the intervals of (α, β) *that are contiguous to* P, *and let* $W(\alpha_n, \beta_n)$ *denote the upper limit of* $|V(\alpha', \beta')|$, *for all intervals* (α', β') *contained in* (α_n, β_n). *Let it be assumed that the series* $\sum\limits_{n=1}^{\infty} W(\alpha_n, \beta_n)$ *is convergent. Then* $V(\alpha, \beta)$ *is defined by*

$$V(\alpha, \beta) = \sum_{n=1}^{\infty} V(\alpha_n, \beta_n) + \int_{(P)} f(x)\, dx.$$

It is clear that $W(\alpha_n, \beta_n)$ is the fluctuation of $V(\alpha_n, x)$ in the interval (α_n, β_n).

It will be shown that the D-integral of $f(x)$, in (a, b), exists, or in accordance with the expression employed by Denjoy, that $f(x)$ is totalizable in (a, b), provided $f(x)$ satisfies the following conditions:

I. *For every perfect set* P, *in* (a, b), *the set of those points of* P *which are points of non-summability of* $f(x)$, *with respect to* P, *is non-dense in* P.

In particular, if P consists of all the points of (a, b), the set H, of points of non-summability of $f(x)$, with respect to the interval (a, b), is a non-dense set.

II. *If* (α, β) *be any interval in* (a, b), *and* $V(\alpha', \beta')$ *has been calculated for every interval* (α', β'), *interior to* (α, β), *then* $V(\alpha', \beta')$ *tends to a definite limit, as* α', β' *converge independently to* α, β *respectively.*

When this condition is satisfied $V(\alpha, \beta)$ has been defined, in (3) above, to be the value of this limit, thus

$$\int_{\alpha}^{\beta} f(x)\, dx = \lim_{\alpha' \sim \alpha, \beta' \sim \beta} \int_{\alpha'}^{\beta'} f(x)\, dx.$$

III. *For every perfect set* P, *if* $V(\alpha_n, \beta_n)$ *has been calculated for every interval* (α_n, β_n), *contiguous to* P, *the set of points of* P *which are not points of convergence of* $\sum\limits_{n=1}^{\infty} W(\alpha_n, \beta_n)$ *is non-dense in* P.

The series $\sum\limits_{n=1}^{\infty} W(\alpha_n, \beta_n)$ is said to be convergent *in an interval* (α', β') if the series consisting of those terms for which (α_n, β_n) is interior to (α', β') is convergent. The series is said to be convergent *at a point* p if p is interior to some interval (α', β') in which the series is convergent.

If every point of a perfect set P is such as to be a point of convergence of $\sum\limits_{n=1} W(\alpha_n, \beta_n)$, then the series $\sum\limits_{n=1} W(\alpha_n, \beta_n)$, taken for all the intervals contiguous to P, is convergent.

For, let a system of nets be fitted on to the interval (a, b), in which P is contained; then if $\sum\limits_{n=1} W(\alpha_n, \beta_n)$ is divergent, there must be at least one mesh d_1, of the net D_1, in which $\sum\limits_{n=1} W$ is divergent. Of the nets of D_2, that are contained in d_1, there must be one at least in which $\sum\limits_{n=1} W$ is divergent; let d_2 be that net, or that of lowest rank if there is more than one. Proceeding in this manner, we obtain a sequence of meshes $d_1, d_2, \ldots$, defining a point Q, in all of them, such that, in each of these meshes, the series $\sum_{n=1} W$ is divergent. This point Q is one of divergence of the series, and it must clearly belong to the set P, which by hypothesis contains no such point. Therefore the series $\sum_{n=1} W(\alpha_n, \beta_n)$, taken for all the intervals contiguous to P, is convergent

It is clear that those points of a perfect set P which are points of divergence of $\Sigma W(\alpha_n, \beta_n)$ form a closed set. For, in any neighbourhood of a limiting point of the set, there are points of the set; from which the result follows.

465. It will be shewn that, when the conditions given in § 464 are satisfied, the D-integral of $f(x)$ can be calculated by means of an enumerable set of L-integrals, of passages to the limit, and of summations in a certain order.

Every point of (a, b) which does not belong to H, the set of points of non-summability of $f(x)$ with respect to the interval (a, b), is interior to a contiguous interval of H. Since $f(x)$ has an L-integral in any interval (α', β'), interior to (α, β), one of the intervals contiguous to H, $V(\alpha, \beta)$ can be determined as $\lim\limits_{\alpha'\sim\alpha,\, \beta'\sim\beta} V(\alpha', \beta')$, in accordance with the condition II, of § 464. Thus the D-integral is determined for each interval contiguous to H. Now H is the sum of a perfect set P_1, the nucleus of H, and an enumerable set M_1; we proceed to calculate $V(\alpha, \beta)$ in the intervals (α, β) that are contiguous to P_1.

The part of M_1 in such an interval (α, β) is reducible, and therefore has a derivative of order $\bar{\gamma}$, some number of the first, or of the second, class, which contains no points, whereas the derivatives of lower order than $\bar{\gamma}$ all exist. If an interval (α', β'), in (α, β), contains no point of M_1, the D-integral in (α', β') is an L-integral; and by condition II the integral can be calculated for (α', β'), when α', β' are points of M_1, whereas no interior points belong to M_1. If (α', β') contains a finite set of points of M_1, the D-integral, for (α', β'), can then be calculated by means of the definition (2) of § 464. Next, let (α_1, β_1) be an interval contiguous to the first derivative M_1', of M_1; then in an interval (α_1', β_1') interior to (α_1, β_1) there is only a finite set of points of M_1; hence, as before, the D-integral in (α_1, β_1) can be calculated. The integrals in all the intervals contiguous to M_1' having thus been determined, by proceeding as before, the integrals in all the intervals contiguous to M_1'' may be determined. Generally, let γ be a finite number, or a transfinite number of the first species. As before, if the D-integral has been calculated for every interval contiguous to $M_1^{(\gamma-1)}$, it can be calculated for every interval contiguous to $M_1^{(\gamma)}$. If

γ be a transfinite number of the second species, let us suppose that the D-integral is known for every interval contiguous to $M_1^{(\gamma')}$, for all values of $\gamma' < \gamma$. In any interval $(\alpha_\gamma', \beta_\gamma')$ which is interior to an interval $(\alpha_\gamma, \beta_\gamma)$ contiguous to $M_1^{(\gamma)}$, from and after some value of γ', all the sets $M_1^{(\gamma')}$ have no points in $(\alpha_\gamma', \beta_\gamma')$; hence the D-integral can be calculated in $(\alpha_\gamma', \beta_\gamma')$, and consequently, by employing the definition (3), it can be determined in $(\alpha_\gamma, \beta_\gamma)$. It has thus been shewn that the D-integral can be calculated for every interval (α, β), contiguous to $M_1^{(\gamma)}$, where γ is any number of the first, or of the second, class. Since $M_1^{(\bar\gamma)} \equiv 0$, for some number $\bar\gamma$, it follows that the D-integral can be calculated for any interval contiguous to P_1, by means of an enumerable set of additions, and of passages to the limit, in which the L-integrals over intervals in which $f(x)$ is summable are employed.

Let us suppose this process to have been carried out for each interval contiguous to P_1, the nucleus of H. In case $f(x)$ is summable relative to P_1, and if every point of P_1 is a point at which the series $\Sigma W(\alpha_n, \beta_n)$ is convergent, we have

$$\int_a^b f(x)\,dx = \int_{(P_1)} f(x)\,dx + \sum_{n=1}^{\infty} \int_{\alpha_n}^{\beta_n} f(x)\,dx.$$

When P_1 does not satisfy these conditions we proceed to analyse the set P_1. It should be observed that the D-integral is determined over any interval (α, β) such that p_1, the part of P_1 contained within it, satisfies the conditions that $f(x)$ is summable over p_1, and that each point of p_1 is a point at which the series $\Sigma W(\alpha_n, \beta_n)$ is convergent.

In accordance with the conditions I and III, assumed to be satisfied, the set H_2, of points of P_1 which are either points of non-summability of $f(x)$ with respect to P_1, or are points of non-convergence of $\Sigma W(\alpha_n, \beta_n)$, is non-dense in P_1; moreover the set H_2 is closed. Let $H_2 = P_2 + M_2$, where P_2 is a perfect set, the nucleus of H_2, and M_2 is an enumerable set. It will be shewn that the D-integral can be calculated, first, in every interval having no interior points that belong to H_2, and secondly, for every interval contiguous to P_2.

Consider an interval δ, which contains within it, and at its ends, no point of H_2. We need only consider the case when δ contains, within it, points of P_1. The interval δ can be decomposed into a portion p_1, of P_1, and a set of intervals, all contiguous to p_1, except possibly two of them which are semi-contiguous to p_1. The function $f(x)$ is summable in p_1, and it has therefore an L-integral over p_1. The series consisting of those terms of $\Sigma W(\alpha_n, \beta_n)$ that refer to intervals contained in δ is convergent. The D-integral over δ is therefore the sum of the D-integrals in the intervals contiguous and semi-contiguous to p_1, together with the L-integral over the set p_1, in accordance with definition (4). By passing to the limit, the

D-integral is obtained over any interval which has no point of H_2 in its interior, but of which one extremity, or both extremities, belong to H_2. By means of an enumerable set of passages to the limit, the D-integral is thus determined over any interval in which H_2 is reducible, as in the former case. Thus the D-integral is determined for any interval contiguous to P_2.

From P_2, we proceed, as before, to define H_3, which is composed of a perfect set P_3, the nucleus of H_3, and an enumerable set M_3. This set H_3 consists of the points of non-summability of $f(x)$ with respect to P_2, and of points of divergence of ΣW.

In general, we obtain a set $H_\gamma = P_\gamma + M_\gamma$, where γ is a number of the first, or of the second, class, and where it is assumed that $H_{\gamma'}$ and $P_{\gamma'}$ are known for all numbers γ', less than γ. In case γ is of the first species, we suppose the D-integral of $f(x)$ to have been already determined in every interval contiguous to $P_{[\gamma-1]}$; the process of determining the D-integral over every interval contiguous to P_γ is the same as in the special case $\gamma = 2$, which has been considered above. If γ is a number of the second species, P_γ is the set of points common to all the sets $P_{\gamma'}$, where γ' has every value $< \gamma$. If none of the sets $P_{\gamma'}$ is devoid of points, the set H_γ certainly exists, but P_γ does not exist in case H_γ is enumerable. The set P_γ, when it exists, is non-dense in each of the sets $P_{\gamma'}$.

It will be shewn that the D-integral can be determined over every interval contiguous to P_γ, it being assumed that the corresponding determination has been made for all the sets $P_{\gamma'}$, where $\gamma' < \gamma$.

First, let γ be a number of the first species; and let δ be an interval which contains, as interior point, or as an end-point, no point of H_γ. If $p_{[\gamma-1]}$ be the portion of $P_{[\gamma-1]}$ in δ, $f(x)$ is summable in $p_{[\gamma-1]}$, and those of the D-integrals which are taken over intervals contiguous to $p_{[\gamma-1]}$ form an absolutely convergent series. We calculate the L-integral of $f(x)$ over $p_{[\gamma-1]}$, and add to it the sum of the D-integrals over the intervals that are contiguous or semi-contiguous to $p_{[\gamma-1]}$; the result is the D-integral over δ. This D-integral is now determined over every interval which contains within it, or at its ends, no point of H_γ. By an enumerable set of passages to the limit the D-integrals are now determined in every interval in which H_γ is reducible, that is, in every interval of which no interior point belongs to P_γ.

Next, let γ be a number of the second species, and let δ have the same meaning as before. The sets $P_{\gamma'}$ $(\gamma' < \gamma)$ have no points in δ, and therefore the D-integral can be calculated over δ. By a passage to the limit, the D-integral can be determined over any interval which has points of H_γ only at one, or both, of the end-points. Hence, by an enumerable set of additions and passages to the limit, the D-integral is calculated over every interval contiguous to P_γ.

All the integrals are obtained by an enumerable set of operations of the types laid down at the beginning of the discussion. Since P_γ is non-dense in all the sets $P_{\gamma'}$, where $\gamma' < \gamma$, there exists a number $\bar{\gamma}$, for which P_γ does not exist (see § 82). Hence, by an enumerable set of operations, the D-integral of $f(x)$ over the interval (a, b) can be calculated, since, in the interval (a, b), there are no points of $P_{\bar{\gamma}}$.

It is clear that the D-integral over (a, x) can be determined in the same manner, where x is any point in (a, b).

466. In the construction of the D-integral, given in § 465, the conditions I and III have been employed only in relation to the particular perfect sets $P_1, P_2, \ldots$, whereas the function $f(x)$, when integrable (D) in (a, b), has been required to be such that these conditions are satisfied in relation to every perfect set P. It will however be shewn that the D-integral constructed by the method indicated in § 465 actually satisfies the conditions I and III, of § 464, in relation to any perfect set P.

In the first place it may be observed that, if $f(x)$ be summable in a sub-interval, and Q be any closed set in that sub-interval, there are no points of Q which are points of non-summability with respect to Q, and also no points at which the series, of which the terms are the fluctuations of $\int_a^x f(x)\,dx$ in the intervals contiguous to Q, is not convergent. If P be any perfect set in (a, b), let K be the set of those points of P such that, at any one of them, either $f(x)$ is not summable with respect to P, or is not a point of convergence of ΣW_n, taken for the intervals contiguous to P. Unless K is non-dense in P, there is some interval in which K and P coincide. Let $P^{(1)}$ be such a portion of P; then $P^{(1)}$ must be contained in H. For, otherwise, a part of $P^{(1)}$ would be contained in an interval interior to an interval contiguous to H; and this is not possible, since $f(x)$ is summable in such an interval. Since the perfect set $P^{(1)}$ is contained in H, it must be contained in P_1. Similarly, if a portion of $P^{(1)}$ were contained in an interval contiguous to H_2, since $P^{(1)}$ is contained in P_1, that portion of $P^{(1)}$ would be a part of P_1 over which $f(x)$ is summable, and such that every point would be one of convergence of ΣW over the intervals contiguous to the part; and this is not possible. Hence $P^{(1)}$ is contained in H_2, and therefore in P_2. Proceeding in this manner, it can be shewn that $P^{(1)}$ is contained in all the sets $P_3, P_4, \ldots P_\alpha$. But P_α, for some value of α, contains no points, and therefore the set $P^{(1)}$ cannot exist. It has thus been shewn that K is non-dense in P. Therefore the conditions I and III are satisfied for any perfect set P.

In the construction given in § 465, the conditions in (3) and II have only been employed when the limiting interval (α, β) is a contiguous interval of one of the sets $H, H_2, \ldots H_{\bar{\gamma}}$. But, as the condition is to be applicable to any interval (α, β), it is necessary to shew that the function V, constructed by the process given in § 465, actually satisfies this condition for every interval (α, β), before the existence of the D-integral can be regarded as completely established. The condition amounts to a postulation that the integral $\int_a^x f(x)\,dx$, or $F(x)$, shall be continuous in the whole interval (a, b). For the continuity of $F(x)$ follows from the definition (3), and the corresponding assumption II. If $x_1, x_2, \dots x_n, \dots$ be a sequence of points interior to (a, x), and converging to x,

$$\lim_{n \sim \infty} V(a, x_n) = V(a, x);$$

and thus $F(x)$ is continuous on the left. Similarly, by considering $V(x, b)$, it may be shewn to be continuous on the right, and therefore $V(a, x)$, which, by definition (2), is $V(a, b) - V(x, b)$, is continuous on the right. Thus the D-integral, $V(a, x)$, when it exists, must be continuous in the interval (a, b).

To prove that the function constructed in § 465 satisfies this condition, we observe, in the first place, that $V(a, x)$ is certainly continuous at any point that does not belong to H. For such a point is interior to an interval in which $f(x)$ is summable; hence $V(a, x+h) - V(a, x)$ is the L-integral $\int_x^{x+h} f(x)\,dx$, which converges to zero, as $h \sim 0$. Next, let G be a perfect set of points, contained in an interval (α, β), and such that $f(x)$ is summable over G, and also such that every point of G is a point of convergence of the series ΣW, when the summation is taken for the intervals, contained in (α, β), which are contiguous to G. We have then,

$$V(a, x+h) - V(a, x) = \left[\int_{(G)} f(x)\,dx + \Sigma \int_{a_n}^{\beta_n} f(x)\,dx\right]_x^{x+h},$$

where (α_n, β_n) is contiguous to G. As $h \sim 0$, $\left|\int_{(G)} f(x)\,dx\right|_x^{x+h}$ converges to zero, on account of a known property of the L-integral In case x and $x+h$ are both points of G, $\left[\Sigma \int_{a_n}^{\beta_n} f(x)\,dx\right]_x^{x+h}$ also converges to zero, as h does so. But if $x+h$ is not a point of G, it is within some interval (α_m, β_m) contiguous to G, and we have to consider a part $\int_{x+h}^{\beta_m} f(x)\,dx$. We have $\left|\int_{a_m}^{x+h} f(x)\,dx\right| \leq W_m$, and therefore, in virtue of the hypothesis made above, this part makes no difference in the convergence of $\left[\Sigma \int_{a_n}^{\beta_n} f(x)\,dx\right]_x^{x+h}$ to zero, when convergence takes place on account of the assumption that x is a point of convergence of ΣW. It has thus been shown that

$$V(a, x+h) - V(a, x)$$

converges to zero, as h converges to zero through any sequence of values; therefore $V(a, x)$ is continuous in (α, β).

If G be identified successively with the perfect sets in the intervals contiguous to $P_1, P_2, \ldots$, the perfect sets employed in § 465, in the construction of $V(a, x)$, we see that it is sufficient for the continuity of that integral that the condition $\lim\limits_{n\sim\infty} W_n = 0$ should be satisfied for all the perfect sets contained in the intervals contiguous to $P_1, P_2, \ldots$. The continuity of $\int_a^x f(x)\,dx$ in the whole interval (a, b) is thus established.

THE FUNDAMENTAL THEOREM OF THE INTEGRAL CALCULUS FOR THE DENJOY INTEGRAL

467. With a view to proving the fundamental theorem that the indefinite Denjoy integral has a finite differential coefficient, equal to the integrand, almost everywhere in the interval of integration, a property of certain types of sets of intervals will be investigated.

It has been shewn, in § 71, that a set Δ of open intervals all contained in the given interval (a, b), can be replaced by a non-overlapping set of open intervals $\overline{\Delta}$, the two sets $\Delta, \overline{\Delta}$ containing the same open set of points.

The set of open intervals Δ is said to be a* *restricted set of intervals*, if no interval of Δ can be removed without altering the equivalent set $\overline{\Delta}$. This is the same as the condition that each interval of Δ contains a point that is not in any of the other intervals of Δ.

It is easily seen that, if three open intervals $\delta_1, \delta_2, \delta_3$ are such that a point P belongs to all of them, one at least of the intervals is such that every point of it belongs to one of the other two. Accordingly, that interval may be removed without affecting the equivalent set of points.

It follows at once that, if Δ be a finite set of intervals, and is a restricted set, any point P is contained in at most two of the intervals of Δ. For such a set of intervals, the sum $\Sigma m(\delta)$, of the measures of all the intervals, cannot exceed $2m(\overline{\Delta})$, or twice the measure of the equivalent set of open intervals.

A set of open intervals Δ is said to be *complete*, if it satisfies the condition that, when $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots (\alpha_n, \beta_n), \ldots$ is any sequence of intervals all belonging to Δ, and such that $\alpha_1, \alpha_2, \ldots \alpha_n, \ldots$ converges to a point α_ω, and $\beta_1, \beta_2, \ldots \beta_n, \ldots$ converges to a point β_ω, then $(\alpha_\omega, \beta_\omega)$ is an interval of Δ. In case $(\alpha_\omega, \beta_\omega)$ is not necessarily an interval of Δ, but is contained in such an interval (α, β), the set is said to be *semi-complete*.

If the set Δ is complete (or semi-complete), a part Δ_1, of Δ, exists which is equivalent to the same non-overlapping set $\overline{\Delta}$, as Δ, and such that it is a restricted set. For, considering the closed set $C(\Delta)$, the complement of $\bar{\Delta}$, let (α, β) be one of its contiguous intervals. Any point P, of the open interval (α, β), is interior to an interval δ, of Δ.

---

* See Denjoy, *Journal de Math* (7), vol. I (1915), p. 223, where the theory here given is developed.

If there exist intervals (α, α'), (β', β) belonging to Δ, the closed interval (α', β') can be covered by a finite set of intervals of Δ, in accordance with the Heine-Borel theorem. This finite set, together with (α, α') and (β', β), forms a part of Δ which covers the open interval (α, β), and this set can then be reduced to a restricted set, by suppressing, if necessary, some of the intervals.

If no such intervals as (α, α'), (β', β) exist, we choose a set of points $(\ldots P_{-n}, \ldots P_{-1}, P_0, P_1, \ldots P_n, \ldots)$, contained in (α, β), such that α is the limiting point of the sequence $(\ldots P_{-n}, \ldots P_{-1}, P_0)$, and that β is the limiting point of the sequence $(P_0, P_1, \ldots P_n, \ldots)$, any point P_{m+1} being taken to be on the right of P_m, whether m be positive or negative.

If the points $\ldots P_{-n}, \ldots P_{-1}, P_0$ are contained respectively in intervals $\ldots \delta_{-n}, \ldots \delta_{-1}, \delta_0$, of Δ, the measure of δ_{-n} must converge to zero, as $n \sim \infty$, for otherwise their right-hand end-points would have a limiting point α'' $(> \alpha)$ and (α, α'') would be an interval of Δ, if the set Δ is restricted, or would be contained in an interval of Δ, in case the set is semi-restricted; and this is contrary to the hypothesis that no interval of Δ has its left-hand end-point at α. A similar statement holds good as regards β.

Every point of the closed interval $P_m P_{m+1}$, where m is positive or negative, is a point of a finite set Δ_m, of intervals of Δ. Let Σ denote the enumerable set of intervals obtained by adding the sets Δ_m, for all values of m. It will be shewn that, if (A, B) is any interval, such that $A > \alpha$, $B < \beta$, the interval (A, B) has points in common with only a finite part of the set Σ. Let $\Sigma(A, B)$ denote the set of those intervals of Σ which contain at least one point of (A, B). Let the intervals of Σ be denoted by $\Sigma_1, \Sigma_2, \ldots \Sigma_\iota, \ldots$, then, if $\Sigma(A, B)$ is not a finite set, it contains intervals Σ with indices which increase indefinitely. There must then be in $\Sigma(A, B)$ intervals of Δ_m, for values of m that increase indefinitely, and are all positive, or else are all negative. In the former case their right-hand end-points will converge to β, and in the latter case their left-hand end-points will converge to α. But it has been shewn that the measure of such an interval, in either case, converges to zero, and thus an infinite set of such intervals cannot all contain points of (A, B). It has thus been shewn that the set $\Sigma(A, B)$ is a finite component of the set Σ.

The case in which one of the intervals (α, α'), (β', β) exists, but not the other, can be reduced to the preceding cases.

The set Σ can be reduced to a restricted set $\bar{\Sigma}$, by suppressing some of the intervals, without altering the set of interior points. To prove this, let ι_1 be the smallest value of ι such that Σ_{ι_1} contains a point that does not belong to Σ_{ι_1+k}, for $k = 1, 2, 3, \ldots$. If P_1 is any point within (α, β), it is interior to only a finite number of the intervals Σ_ι; let N_1 be the greatest index such that P_1 is interior to Σ_{N_1}, we see then that $\iota_1 \leq N_1$. The interval Σ_{ι_1} having thus been determined, we proceed to determine ι_2 as the smallest index $(> \iota_1)$ such that it contains a point that does not belong either to Σ_{ι_2+k}, for $k = 1, 2, 3, \ldots$, or to Σ_{ι_1}. If we take a point P_2, not in Σ_{ι_1}, there is a greatest number N_2 such that P_2 is a point of Σ_{N_2}; thus $\iota_2 \leq N_2$. In general if $\Sigma_{\iota_1}, \Sigma_{\iota_2}, \ldots \Sigma_{\iota_p}$ have been determined $\Sigma_{\iota_{p+1}}$ is determined by the condition that ι_{p+1} is the smallest integer $(> \iota_p)$ such that $\Sigma_{\iota_{p+1}}$ contains a point that does not belong to $\Sigma_{\iota_{p+1}+k}$, for $k = 1, 2, 3, \ldots$, and that does not belong to any of the intervals $\Sigma_{\iota_1}, \Sigma_{\iota_2}, \ldots \Sigma_{\iota_p}$. If P_{p-1} is a point that does not belong to any of the intervals $\Sigma_{\iota_1}, \Sigma_{\iota_2}, \ldots \Sigma_{\iota_p}$; and if N_{p+1} is the greatest of the indices $\iota_{p+1}, \iota_{p+2}, \ldots$ such that $\Sigma_{N_{p+1}}$ contains P_{p+1}, we have then $\iota_{p+1} \leq N_{p+1}$. The set of intervals $\{\Sigma_{\iota_p}\}$ is a restricted set, and it can be shewn to cover the whole of the open interval (α, β). For, let Q be any point of that interval, then there exists a finite set $\Sigma_{\gamma_1}, \Sigma_{\gamma_2}, \ldots \Sigma_{\gamma_m}$, of intervals of Σ, all of which contain Q. If none of the indices $\gamma_1, \gamma_2, \ldots \gamma_{m-1}$ belong to the set $\{\iota_q\}$, it will be shewn that γ_m must belong to that set, and thus that Σ_{γ_m} is one of the intervals of the restricted set Σ. The interval Σ_{γ_m} contains a point that does not belong to any of the intervals $\Sigma_{\iota_1}, \Sigma_{\iota_2}, \ldots \Sigma_{\iota_p}$, nor to Σ_{ι_p+h}, for $h = 1, 2, 3, \ldots$; and therefore, for a properly chosen value of p, it must have the least index for which these conditions are satisfied, and thus $\gamma_m = \iota_{p+1}$.

If we consider all the intervals (α, β) contiguous to $C(\overline{\Delta})$, we obtain a restricted set of intervals that has the same interior points as the given set Δ. Moreover, a restricted set of intervals which covers the open interval (a, b), has only a finite component such that each interval of that component contains a point of a given interval (A, B), interior to (a, b). It has been shewn that the measure of (A, B) is not less than half the sum of the lengths of the intervals of this finite set. By considering a sequence of intervals (A, B) converging to (a, b), we see that the *sum of the lengths of the intervals of a restricted set which covers the open interval* (a, b) *cannot exceed twice the length* $b - a$.

468. The properties of a restricted set of intervals, covering an open segment, will now be applied to prove the following theorem:

Let P be a perfect set, contained in the linear interval (a, b), and the intervals $u_1, u_2, \ldots u_n, \ldots$ be contiguous to P, and let $\chi_n(x)$ be defined in each interval u_n, or (a_n, b_n), and such that $\chi_n(a_n) = 0$. Let $w(u_n)$ or w_n denote the upper boundary of $|\chi_n(x)|$ in (a_n, b_n), and let it be assumed that the series $\sum\limits_{n=1} w(u_n)$ is convergent. Let the function $F(x) = \sum\limits_a^x \chi_n(b_n) + \chi_m(x)$

consist of those terms for which $b_n < x$, and, in case x is interior to one of the intervals, (a_m, b_m), of the term $\chi_m(x)$, for the portion (a_m, x) of the interval; then $F(x)$ is such that it has a differential coefficient equal to zero, at almost all points of P.

Let the intervals $\rho_1, \rho_2, \dots \rho_{N+1}$ be complementary to the set of N intervals $u_1, u_2, \dots u_N$.

Let β denote any segment of which the ends are points of P, and let $\sigma(\beta)$ denote the sum of $|\chi_n(b_n)|$, for all those intervals (a_n, b_n) that are in β.

Let A, N be two positive numbers, and let $\delta(N, A)$ be an interval, such that its end-points belong to P, and that it is contained in one of the intervals $\rho_1, \rho_2, \dots \rho_{N+1}$, and is also such that $\sigma\{\delta(N, A)\}$ is not less than $A^{-1}m\{\delta(N, A)\}$.

Let $\Delta(N, A)$ be the set of all such intervals $\delta(N, A)$; it will be shewn that this set of intervals is a complete set. If (α_p, β_p) be an interval of a sequence, all the intervals of which belong to $\Delta(N, A)$, and such that α_p, β_p converge to α, β respectively, as $p \sim \infty$; it will be seen that (α, β) is an interval of the set $\Delta(N, A)$. In the first place, since α_p, β_p are points of P, for every value of p, it follows that α, β are points of P. Moreover, from and after some fixed value of p, all the intervals (α_p, β_p) must be in one and the same interval ρ_m; hence (α, β) is also in that interval. The smallest index n, of an interval u_n contained in (α, α_p), or in (β, β_p), must increase indefinitely, as p does so, thus $\sigma(\alpha_p, \alpha)$, $\sigma(\beta, \beta_p)$ converge to zero, as $p \sim \infty$; hence

$$\frac{1}{\beta - \alpha}\sigma(\alpha, \beta) = \lim_{p\sim\infty}\frac{1}{\beta_p - \alpha_p}\sigma(\alpha_p, \beta_p) \geq \frac{1}{A}.$$

It has thus been shewn that (α, β) satisfies all the necessary conditions that it may belong to $\Delta(N, A)$. The set $\Delta(N, A)$ being a complete set of intervals, it follows that it can be replaced by a restricted set of intervals $\overline{\Delta}(N, A)$, containing the same interior points. The points interior to at least one interval of $\Delta(N, A)$, that is, to at least one interval of $\overline{\Delta}(N, A)$, form a non-overlapping set $\iota(N, A)$ of open intervals. Every point of one such interval ι is interior to at least one, and to not more than two, intervals of $\overline{\Delta}(N, A)$. Thus the end-points of the intervals of $\overline{\Delta}(N, A)$, that are interior to ι, divide it into segments alternately in one, and in two, of the intervals of $\overline{\Delta}(N, A)$; let $\overline{\Delta}_1(\iota)$ denote the first, and $\overline{\Delta}_2(\iota)$ the second, set of these segments. We have then

$$m(\iota) = m\{\overline{\Delta}_1(\iota)\} + m\{\overline{\Delta}_2(\iota)\},$$

and also

$$\sigma(\iota) = \sigma\{\overline{\Delta}_1(\iota)\} + \sigma\{\overline{\Delta}_2(\iota)\}.$$

Every interval $\overline{\delta}(N, A)$ having points in common with ι, and consequently contained in ι, is decomposed into one interval of $\overline{\Delta}_1(\iota)$ and two of $\overline{\Delta}_2(\iota)$; hence we have

$$\sigma\{\overline{\Delta}(\iota)\} = \sigma\{\overline{\Delta}_1(\iota)\} + 2\sigma\{\overline{\Delta}_2(\iota)\}.$$

Now $$\sigma\{\bar{\Delta}(\iota)\} \geq A^{-1}m\{\bar{\Delta}(\iota)\} \geqq A^{-1}m(\iota),$$
and $$\sigma\{\bar{\Delta}_1(\iota)\} + 2\sigma\{\bar{\Delta}_2(\iota)\} < 2\sigma(\iota);$$
therefore $m(\iota) < 2A\sigma(\iota)$. This inequality holds for each interval covered by $\Delta(N, A)$, and these intervals ι are interior to the segments
$$\rho_1, \rho_2, \dots \rho_{N+1}.$$
The intervals u_n, interior to an interval ι, all have indices greater than N, hence, denoting by t_N the sum $\sum_{N+1}^{\infty} |\chi_n(b_n)|$, the total measure of all the intervals ι, covered by the intervals $\Delta(N, A)$, is $< 2At_N$. This measure diminishes indefinitely as N is increased indefinitely.

Let each interval u_n be increased at each end by adding on a segment of length Aw_n; and denote the interval so increased by $u_n'(A)$.

If we exclude from (a, b) all the segments $u_1, u_2, \dots u_N$, and the segments $\Delta(N, A)$, and also the segments $u'_{N+p}(A)$, where $p = 1, 2, 3, \dots$, the measure of these excluded segments is less than
$$\sum_1^N u_n + 2At_N + \sum_{N+1}^{\infty} u_n'(A),$$
or than $$\sum_1^{\infty} u_n + 4As_N, \text{ where } s_N = \sum_{N+1}^{\infty} w_n.$$

The measure of P being $b - a - \sum_1^{\infty} u_n$, the measure of the points excluded is at most $b - a - m(P) + 4As_N$. If N is so large that $4As_N < m(P)$, the interval (a, b) contains points that are not excluded. These points form a set $E(N, A)$ of which the measure is at least $m(P) - 4As_N$; and this set is contained in P, since it does not contain any points of u_n, for any value of n. Moreover $E(N, A)$ is contained in $E(N+1, A)$, as is seen by observing that every interval of $\Delta(N+1, A)$ is an interval of $\Delta(N, A)$. If $E(A)$ denote the outer limiting set of the sequence of sets $\{E(N, A)\}$, as N is increased indefinitely, we see that $E(A)$ is contained in P, and has the same measure as P, since $4As_N \sim 0$, as $N \sim \infty$.

Let ξ be any point of $E(A)$; it then belongs to $E(N, A)$ for a certain value of N, and for all greater values; it is moreover a point of P which is a limiting point of P, on both sides. The point ξ is interior to a segment ρ_ι belonging to the finite set $\rho_1, \rho_2, \dots \rho_{N+1}$; a sequence of values of x converging to ξ may be taken to be interior to ρ_ι.

If all the points x, of the sequence, are points of P, we have
$$|F(x) - F(\xi)| \leqq \sum_{\xi}^{x} |\chi_n(b_n)| = \sigma(\xi, x), \text{ if } x > \xi,$$
and $$|F(x) - F(\xi)| \leqq \sigma(x, \xi), \text{ if } x < \xi.$$

Since ξ is in the set $E(N, A)$, we see that $\sigma(\xi, x)$, or $\sigma(x, \xi)$, is $< A^{-1}|x - \xi|$; hence $\left|\dfrac{F(x) - F(\xi)}{x - \xi}\right| < \dfrac{1}{A}$.

If x be a point of ρ_ι that does not belong to P, and is thus interior to u_m, for some value of m, greater than $N+1$, we have $|x-\xi| > Aw_m$, since ξ is not in the closed interval $u'_m(A)$.

If $x > \xi$, we have $F(x) - F(\xi) = \chi_m(x) + F(a_m) - F(\xi)$: hence

$$|F(x) - F(\xi)| < w_m + \frac{1}{A}|a_m - \xi| < \frac{2}{A}|x - \xi|.$$

If $x < \xi$, we have

$$F(\xi) - F(x) = F(\xi) - F(b_m) + \chi(b_m) - \chi_m(x);$$

and therefore

$$|F(x) - F(\xi)| < \frac{1}{A}|\xi - b_m| + 2w_m < \frac{3}{A}|\xi - x|.$$

It is now seen that, for any point x, in ρ_ι, $\left|\frac{F(x) - F(\xi)}{x - \xi}\right| < \frac{3}{A}$; hence none of the four derivatives of $F(x)$, at ξ, can exceed $3/A$. If we assign to A, the values in an increasing sequence that diverges, we see that each of the sets $E(A)$ is contained in the preceding set; for $E(N, A')$ is contained in $E(N, A)$, if $A' > A$. Each of the sets $E(A)$ has the measure $m(P)$, and therefore the set E of points common to all the sets $E(A)$ has the measure $m(P)$. At a point ξ, of E, the four derivatives of $F(\xi)$ are all less than $3/A$, for all the values of A in the sequence. Therefore $F(\xi)$ has a differential coefficient, equal to zero, at every point of E, that is, at almost every point of P. The theorem has thus been established. Another proof of this result has been given* by Burkill.

469 From the theorem of § 468, the following theorem can be deduced at once:

If P be a perfect set, and $\{(a_n, b_n)\}$ be the set of contiguous intervals, then, if the function $V(a, x)$ be summable over P, and if $\Sigma W(a_n, b_n)$ be convergent, the function $V(a, x)$ has a finite differential coefficient, at almost all points of P.

The term $W(\alpha, \beta)$ denotes, as in § 464, the upper limit of $|V(\alpha', \beta')|$, for all intervals (α', β') contained in (α, β). In this case

$$\chi_n(x) = \int_{a_n}^{x} f(x)\,dx, \quad \chi_n(b_n) = \int_{a_n}^{b_n} f(x)\,dx.$$

Moreover we have

$$w(a_n, b_n) \leq W(a_n, b_n) \leq 2w(a_n, b_n),$$

and hence, if $\sum_{n=1}^{\infty} W(a_n, b_n)$ is convergent, so also is $\sum_{n=1}^{\infty} w(a_n, b_n)$, and the converse also holds good.

We have $V(a, x) = \int_a^x \bar{f}(x)\,dx + \sum_a^x V(a_n, b_n) + V(a_m, x)$, where the summation is taken for all the intervals (a_n, b_n) interior to (a, x), and the term $V(a_m, x)$ exists if x is interior to (a_m, b_m), and the function $\bar{f}(x)$ is equal to $f(x)$ at every point of P, but is elsewhere zero. Since $\bar{f}(x)$ is summable in (a, b), its indefinite integral has, almost everywhere, a differential coefficient, equal to $\bar{f}(x)$; and by the theorem of § 468, the function $\sum_a^x V(a_n, b_n) + V(a_m, x)$ has a differential coefficient equal to zero, at almost all points of P. It follows that the function $V(a, x)$ has a differential coefficient, equal to $f(x)$, at almost all points of P.

---

* *Proc. Camb. Phil. Soc.* vol. XXI (1923), p. 660.

470. The theorem of § 469 may now be employed to establish the following property of the D-integral:

The indefinite D-integral $F(x)$, of a function $f(x)$, in the interval (a, b), has almost everywhere in the interval a differential coefficient, equal to $f(x)$.

This is the extension to the D-integral, of the property of the L-integral, given in § 405.

With the notation of § 465, it is clear that, in any interval contiguous to H, $F(x)$ has almost everywhere a differential coefficient, equal to $f(x)$. For $f(x)$ has an L-integral in every interval (α', β'), interior to an interval (α, β), contiguous to H, and therefore (§ 405) $F(x)$ has a differential coefficient, equal to $f(x)$, almost everywhere in (α', β') and therefore almost everywhere in (α, β). It can now be shewn that $F(x)$ has this property in any interval contiguous to P_1. For it holds in any interval contiguous to M_1', and therefore also in any interval contiguous to M_1''. Generally, it can be shewn that $F(x)$ has the property in every interval contiguous to any derivative $M_1^{(\gamma)}$, of M_1, and thence it is shewn that $F(x)$ has the property in any interval contiguous to P_1.

It will next be shewn that, in any interval contiguous to P_2, the function $F(x)$ has almost everywhere a differential coefficient equal to $f(x)$. For, consider an interval δ, which contains no points of H_2, but contains a portion p_1, of P_1. The property holds in each interval contained in δ, contiguous to p_1, and by the theorem of § 469, it holds for the set p_1; and consequently it holds for δ. It can now be shewn, as in the case of P_1, that the theorem holds for every interval contiguous to P_2. Proceeding in this manner, as in § 465, we see by induction, that, in virtue of the theorem of § 469, the theorem holds for any interval contiguous to P_γ, where γ is a number of the first, or of the second, class. Since, for some value $\bar{\gamma}$, of γ, the set $P_{\bar{\gamma}}$ is non-existent, it follows that the function $F(x)$ has a differential coefficient, equal to $f(x)$, almost everywhere in (a, b).

471. The importance of the D-integral in the general theory of integration depends largely upon the property of the integral expressed in the following theorem:

If $F(x)$ have, at each point of the interval (a, b), a finite differential coefficient $f(x)$, the function $f(x)$ is integrable (D), in the interval (a, b), and

$$\int_a^x f(x)\,dx = F(x) - F(a).$$

This theorem expresses the fact that the operations of differentiation and integration are completely reversible in the case of any function which has, everywhere in the interval in which it is defined, a finite differential coefficient, whether that differential coefficient is bounded or unbounded in the interval. The property (B), in the statement in § 343, of the fundamental theorem of the differential calculus, is covered, in a wide class of cases, by the above theorem. It will be observed that the corresponding theorem for the L-integral is subject to the condition that $F(x)$ must be of bounded variation, as otherwise $f(x)$ will not be summable (see § 405), whereas, in the above theorem, the function $F(x)$ is subject to no condition beyond that of possessing everywhere a finite differential coefficient.

In order to prove the theorem, it is necessary to shew that the function $f(x)$ satisfies, in relation to $F(x)$, the conditions I, II, and III, of § 461.

(1). To shew that the set of points of non-summability of $f(x)$, with respect to any perfect set P, contained in (a, b), is non dense in P, we observe (see § 401) that $f(x)$ is the limit of a sequence $\{\psi_n(x)\}$ of continuous functions. For, if $\psi_n(x)$ denote $\{F(x + h_n) - F(x)\}/h_n$, where $\{h_n\}$ is a sequence of numbers which converges to zero, and is independent of x, these functions $\psi_n(x)$ are all continuous functions of x.

The function $\psi_n(x)$ may be expressed by $\chi(x, y)$, a function of the two variables x, y, where $y = 1/n$. The function $\chi(x, y)$ is in the first instance defined only for values of y, of the form $1/n$, but it may be extended to the case in which y has all values in the interval $0 < y \leq 1$, by such a rule as that, when y is in the interval $\left(\frac{1}{n+1}, \frac{1}{n}\right)$,

$$\chi(x, y) = \chi\left(x, \frac{1}{n}\right) + \frac{\frac{1}{n} - y}{\frac{1}{n} - \frac{1}{n+1}}\left\{\chi\left(x, \frac{1}{n+1}\right) - \chi\left(x, \frac{1}{n}\right)\right\}.$$

The function $\chi(x, y)$, so defined for the plane set of points $a \leq x \leq b$, $0 < y \leq 1$, is continuous with respect to y, for each value of x. In accordance with the theorem of § 323, there must be, in every interval on the line $y = 0$, points at which $\chi(x, y)$ is continuous with respect to (x, y), and therefore with respect to x. Therefore $f(x)$, which is $\lim_{y=0} \chi(x, y)$, is point-wise discontinuous; and this is also the case if only those values of x are considered which belong to a perfect set P. It follows that the points of infinite discontinuity of $f(x)$, with respect to P, form a set that is non-dense in P. This set contains the set of points of non-summability of $f(x)$ with respect to P. Therefore $f(x)$ satisfies the condition I. of § 464.

(II). If (α, β) is an interval in which $f(x)$ is summable, we have, in accordance with the theorem of § 406, $F(\beta) - F(\alpha) = \int_\alpha^\beta f(x)\,dx$. Let it be now assumed that (α, β) is an interval such that the relation

$$F(\beta') - F(\alpha') = \int_{\alpha'}^{\beta'} f(x)\,dx$$

has been shewn to hold, at some stage of the process of construction of the D-integral of $f(x)$, for all intervals (α', β') interior to (α, β). If $\alpha' \sim \alpha$, $\beta' \sim \beta$, $F(\alpha')$ converges to $F(\alpha)$, and $F(\beta')$ converges to $F(\beta)$, on account of the continuity of the function $F(x)$. It follows that, $\int_\alpha^\beta f(x)\,dx$, defined as $\lim\limits_{\alpha' \sim \alpha,\, \beta' \sim \beta} \int_{\alpha'}^{\beta'} f(x)\,dx$, is equal to $F(\beta) - F(\alpha)$. Therefore the condition II. of § 464, is satisfied.

(III) Let it be assumed that, in each interval (α, β) contiguous to a perfect set P, the D-integral of $f(x)$ has been shewn to exist, and to be equal to $F(\beta) - F(\alpha)$ The following theorem will be proved:

If the continuous function $F(x)$ have, in the perfect set P, a finite differential coefficient at every point, then the set of those points of P which are points of divergence of the series ΣW_n, where W_n is the fluctuation of $F(x)$ in the interval (α_n, β_n) contiguous to P, is non-dense in P

In case the set of points of divergence of the series ΣW_n is not non-dense in P, there must be an interval containing a portion of P, in which the set is everywhere dense in that portion of P. Let it therefore be assumed that the set of points of divergence of ΣW_n is everywhere dense in P. It will then be proved that there exist points of P at which one of the derivatives of $F(x)$ is infinite, and this is inconsistent with the assumption that $F(x)$ has a finite differential coefficient at every point of P. The theorem will then have been established.

In every interval containing a point of P, the series ΣW_n is divergent: it then follows that in this interval there is an infinite number of values of n for which $W_n/(\beta_n - \alpha_n)$ exceeds a fixed positive number k, for if

$$W_n \leqq k(\beta_n - \alpha_n),$$

for all values of n with the exception of a finite number, the sum $\sum\limits_m^\infty W_n$, for some fixed value of m, would be $\leqq k\Sigma(\beta_n - \alpha_n)$, and thus the series would be convergent.

If ξ, η be the points of (α_n, β_n) at which $F(x)$ has its upper and lower limits, we have

$$F(\xi) - F(\eta) = \{F(\xi) - F(\alpha_n)\} + \{F(\alpha_n) - F(\eta)\};$$

and since both the terms on the right-hand side are ≥ 0, one of them is greater than $\frac{1}{2}\{F(\xi) - F(\eta)\}$. There exists therefore a point γ_n, in (α_n, β_n), such that $\left|\frac{F(\gamma_n) - F(\alpha_n)}{\gamma_n - \alpha_n}\right|$ is greater than $\frac{1}{2}\frac{F(\xi) - F(\eta)}{\beta_n - \alpha_n}$. It thus appears that there exists a set of points $c_1, c_2, \ldots c_m, \ldots$ all belonging to P, and dense in P, such that there is a point γ_m on the right of each point c_m, for which $\left|\frac{F(\gamma_m) - F(c_m)}{\gamma_m - c_m}\right| > \lambda$, where λ is an arbitrarily chosen positive number. From the continuity of $F(x)$ at c_m, an interval σ_m containing c_m, of length not exceeding ϵ_m, and not containing the point γ_m, can be so determined that $\left|\frac{F(\gamma_m) - F(x)}{\gamma_m - x}\right|$ lies between the two numbers $\left|\frac{F(\gamma_m) - F(c_m)}{\gamma_m - c_m}\right| \pm \epsilon_m$, for all points x in σ_m. The number ϵ_m is taken to belong to a sequence $\{\epsilon_m\}$, of decreasing numbers, which converges to zero. Any point ξ, contained in an infinite number of the intervals σ_m, belongs to P. For the length of σ_m, not exceeding ϵ_m, which contains the point c_m, of P, the point ξ is a limiting point of P, and therefore belongs to the set. Also $\gamma_m - \xi$ does not exceed $\beta_m - \alpha_m + \epsilon_m$, and therefore γ_m converges to ξ, as m increases indefinitely, it follows that $\left|\frac{F(\gamma_m) - F(\xi)}{\gamma_m - \xi}\right|$, as $m \sim \infty$, has its lower limit $> \lambda$; and thus, at the point ξ, there is a derivative numerically $> \lambda$.

It has now been shewn that all the points ξ, interior to an infinite number of the intervals c_m, form a residual set L relative to the perfect set P. At every such point ξ there is a derivative numerically greater than k.

Let λ have successively the values in an increasing sequence $\{\lambda_n\}$ which diverges, and let $\{L_n\}$ be the corresponding sequence of sets L, residuals with respect to P. A point that does not belong to any of the sets L_n belongs to the sets, of the first category relative to P, that are complementary to the sets L_n, and all these complementary sets form an enumerable sequence of sets, non-dense in P; therefore the points that belong to all the sets $\{L_n\}$ form a residual set, relatively to P. At a point of this residual set, there is a derivative that is numerically greater than λ_n, whatever value n may have, and thus there is an infinite derivative. It has now been shewn that the assumption made is inconsistent with the existence of a finite differential coefficient at each point of P. Thus it has been shewn that the condition III, of § 464, is satisfied.

It has been shewn that the conditions I, II, III, of § 464, are all satisfied by $f(x)$, and it only remains to be proved that the condition (4), of § 464, is satisfied. We shall prove the following theorem:

If $F(x)$ have a finite differential coefficient $f(x)$, at every point of a perfect set P, which is summable over P, and the series $\sum_{n=1}^{\infty} W(a_n, b_n)$ is convergent, where $W(a_n, b_n)$ denotes the fluctuation of $F(x)$ in the interval (a_n, b_n) contiguous to P, then

$$F(b) - F(a) = \int_{(P)} f(x)\,dx + \sum_{n=1}^{\infty} \{F(b_n) - F(a_n)\},$$

where (a, b) is the interval in which P is contained.

In the theorem of § 468, let

$$\chi_n(x) = F(x) - F(a_n),$$

then the function

$$g(x) = F(a) + \sum_a^x \{F(b_n) - F(a_n)\} + \{F(x) - F(a_m)\}$$

has, at almost all points of P, the differential coefficient zero. The summation in the second term on the right-hand side is taken for all the intervals (a_n, b_n), interior to (a, x), and x is taken to be in the interval (a_m, b_m).

Let $F(x) = g(x) + h(x)$, then $h(x)$ has a differential coefficient, equal to $f(x)$, at almost all points of P.

Thus, if $f(x)$ is summable over P, we have

$$F(x) = g(x) + \int_a^x f(x)\,dx,$$

where $\bar{f}(x) = f(x)$ at all points of P, and is elsewhere equal to zero.

It now follows that

$$F(b) - F(a) = \int_{(P)} f(x)\,dx + \sum_{n=1}^{\infty} \{F(b_n) - F(a_n)\}.$$

It has thus been shewn that all the conditions necessary for the existence of the D-integral in (a, b) are satisfied, and that the integral is equal to $F(b) - F(a)$. It is clear that, in the interval (a, x), the D-integral $\int_a^x f(x)\,dx$ exists, and is equal to $F(x) - F(a)$.

PROPERTIES OF THE DENJOY INTEGRAL

472. A D-integral that is not also an L-integral cannot be an absolutely convergent integral, that is, the integral of the absolute value o the function cannot exist. For, if α and β are two points of non-summability of the function $f(x)$, and the L-integral of $f(x)$ exists in every interval (α', β') interior to (α, β), the limit of $\int_{\alpha'}^{\beta'} |f(x)|\,dx$, as $\alpha' \sim \alpha$, $\beta' \sim \beta$, is infinite.

If $f^{(1)}(x)$, $f^{(2)}(x)$ be two functions, both integrable (D), in the interval (a, b), then their sum is also integrable (D), in the interval, and

$$\int_a^b \{f^{(1)}(x) + f^{(2)}(x)\}\,dx = \int_a^b f^{(1)}(x)\,dx + \int_a^b f^{(2)}(x)\,dx.$$

Let $H^{(1)}$, $H^{(2)}$ be the sets of points of non-summability of the two functions $f^{(1)}(x)$, $f^{(2)}(x)$, respectively. The points of non-summability of $f^{(1)}(x) + f^{(2)}(x)$ form the whole, or a closed part, of the closed set $M(H^{(1)}, H^{(2)})$, which may be denoted by $\overline{H}$; and this is the sum of a perfect set $\overline{P}_1$, and an enumerable set $\overline{M}_1$. It has been shewn in § 466, that the set of points of non-summability of either of the functions $f^{(1)}(x)$, $f^{(2)}(x)$, assumed to be integrable (D), with respect to any perfect set whatever, contained in (a, b), is non-dense in that set, and that the same is true as regards the points of non-convergence of the sum of the fluctuations of their integrals in the intervals contiguous to the perfect set. We may therefore, in the process of construction of each of the integrals

$$\int_a^b f^{(1)}(x)\,dx, \quad \int_a^b f^{(2)}(x)\,dx,$$

employ the set $\overline{H}$, instead of $H^{(1)}$, or $H^{(2)}$. We shall have

$$\overline{P}_1 = M(P_1^{(1)}, P_1^{(2)}),$$

and in general $\overline{P}_\alpha = M(P_\alpha^{(1)}, P_\alpha^{(2)})$. Moreover if, at any stage of the process, it has been shown that the theorem holds for every interval (α', β') interior to the interval (α, β), we see that, since the integrals of $f^{(1)}(x)$, $f^{(2)}(x)$ in (α', β') converge to their integrals in (α, β), as $\alpha' \sim \alpha$, $\beta' \sim \beta$, the theorem also holds for the interval (α, β) The same theorem therefore holds for the integrals obtained at each stage of the process of construction of the D-integral of $f^{(1)}(x) + f^{(2)}(x)$, and therefore for the final integrals taken over the interval (a, b).

473. In the case in which the set H, of points of non-summability of the function $f(x)$, has measure zero, the function is necessarily summable over the nucleus P, of H, because $m(P) = 0$, and thus $\int_{(P)} f(x)\,dx = 0$. In this case, the D-integral over (a, b), assumed to exist, is the sum of the D-integrals over the intervals contiguous to P. In order that the D-integral may exist, it is sufficient that the sum of the fluctuations of the integrals over the intervals contiguous to P should be convergent. When this condition is satisfied, so that the D-integral of $f(x)$, over (a, b), exists, it is not necessarily the case that $f(x)$ has an HL-integral over (a, b), because the condition for the existence of the latter is, in accordance with E. H. Moore's theorem, given in § 457, that the sum of the fluctuations of the integrals of $f(x)$ taken over the intervals contiguous to the set H should be convergent; and this condition is more stringent than the condition that the sum of the fluctuations over the intervals contiguous to the nucleus P, of H, should be convergent. This explains why the sum of two functions, both of which have HL-integrals, does not necessarily possess an HL-integral (see § 460), although, as has been shown above, the sum of two functions, each of which is integrable (D), is also integrable (D).

474. The following property of a function that is integrable (D) will now be established:

If $f(x)$ be integrable (D), in the interval (a, b), and $\phi(x)$ be bounded and monotone in the interval, then the product $f(x)\phi(x)$ is integrable (D), in (a, b), and satisfies the relation

$$\int_a^b f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_a^b - \int_a^b F(x)\,d\phi(x),$$

where $F(x)$ denotes the indefinite integral $\int_a^x f(x)\,dx$.

It follows from this theorem that the product of $f(x)$ into any function of bounded variation in (a, b) is integrable (D).

To prove the theorem it is necessary to shew that the function $f(x)\,\phi(x)$ satisfies the conditions given in § 464, in relation to the function

$$\Big[F(x)\,\phi(x)\Big]_a^x - \int_a^x F(x)\,d\phi(x).$$

If, in the interval (α, β), the function $f(x)$ is summable, we have (§ 447)

$$\int_\alpha^\beta f(x)\,\phi(x)\,dx = \int_\alpha^\beta \phi(x)\,dF(x),$$

and therefore

$$\int_\alpha^\beta f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F(x)\,d\phi(x);$$

thus the theorem holds in the interval (α, β).

Next, if the theorem holds in every interval (α', β'), interior to the interval (α, β), it will be shewn to hold for (α, β).

We have to shew that, as $\alpha' \sim \alpha$, $\beta' \sim \beta$,

$$\Big[F(x)\,\phi(x)\Big]_{\alpha'}^{\beta'} - \int_{\alpha'}^{\beta'} F(x)\,d\phi(x)$$

converges to

$$\Big[F(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F(x)\,d\phi(x).$$

Since $\phi(x)$ is monotone, it is clear that $\int_\alpha^{\alpha'} F(x)\,d\phi(x)$ is equal to

$$F(\alpha'')\int_\alpha^{\alpha'} d\phi(x),$$

where α'' is some number in the interval (α, α'), and this is equal to

$$F(\alpha'')\,\{\phi(\alpha') - \phi(\alpha)\};$$

therefore

$$\lim_{\alpha' \sim \alpha}\int_\alpha^{\alpha'} F(x)\,d\phi(x) = F(\alpha)\,\{\phi(\alpha + 0) - \phi(\alpha)\}.$$

Similarly, we have

$$\lim_{\beta' \sim \beta}\int_{\beta'}^{\beta} F(x)\,d\phi(x) = F(\beta)\,\{\phi(\beta) - \phi(\beta - 0)\};$$

and therefore
$$\lim_{\alpha'\sim\alpha,\,\beta'\sim\beta}\int_{\alpha'}^{\beta'} F(x)\,d\phi(x) = \int_{\alpha}^{\beta} F(x)\,d\phi(x)$$
$$+ F(\beta)\{\phi(\beta) - \phi(\beta - 0)\}$$
$$+ F(\alpha)\{\phi(\alpha + 0) - \phi(\alpha)\}.$$

Since
$$\lim_{\alpha'\sim\alpha,\,\beta'\sim\beta}\Big[F(x)\,\phi(x)\Big]_{\alpha'}^{\beta'} = \Big[F(x)\,\phi(x)\Big]_{\alpha}^{\beta} - F(\beta)\{\phi(\beta) - \phi(\beta - 0)\}$$
$$- F(\alpha)\{\phi(\alpha + 0) - \phi(\alpha)\},$$
the result follows.

Hence, as $\int_{\alpha}^{\beta} f(x)\,\phi(x)\,dx$ is defined to be
$$\lim_{\alpha'\sim\alpha,\,\beta'\sim\beta}\int_{\alpha'}^{\beta'} f(x)\,\phi(x)\,dx,$$
we see that the theorem holds for the interval (α, β).

A point in the neighbourhood of which $f(x)$ is summable is also a point in the neighbourhood of which $f(x)\,\phi(x)$ is summable, thus the set H is the same for the function $f(x)\,\phi(x)$, as for $f(x)$, unless $\phi(x) = 0$, in a line of invariability.

We have now to shew that the condition III, of § 464, is satisfied for the function $f(x)\,\phi(x)$.

If (α_n, β_n) be an interval contiguous to the perfect set P, in which the integral of $f(x)\,\phi(x)$ has already been constructed, we have
$$\int_{\alpha_n}^{x} f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{x} - \int_{\alpha_n}^{x} F(x)\,d\phi(x).$$

If ξ, η are the points of the interval at which $F(x)$ attains its upper and lower boundaries, we have to estimate $W'(\alpha_n, \beta_n)$, the fluctuation of
$$\int_{\alpha_n}^{x} f(x)\,\phi(x)\,dx$$
in the interval. The fluctuation of $\Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{x}$ is not greater than
$$\{F(\xi) - F(\alpha_n)\}\,\phi(\beta_n) - \{F(\eta) - F(\alpha_n)\}\,\phi(\alpha_n);$$
it being assumed that $\phi(x)$ is a non-diminishing function. Since
$$F(\xi) - F(\alpha_n) > 0, \quad F(\eta) - F(\alpha_n) < 0,$$
the fluctuation is not greater than $\{F(\xi) - F(\eta)\}\,\phi(\beta_n)$, or than
$$\phi(b)\,W(\alpha_n, \beta_n),$$
where $W(\alpha_n, \beta_n)$ is the fluctuation of $F(x)$.

Since $\int_{\alpha_n}^{x} F(x)\,d\phi(x)$ lies between $F(\xi)\,\{\phi(\beta_n) - \phi(\alpha_n)\}$ and
$$F(\eta)\,\{\phi(\beta_n) - \phi(\alpha_n)\},$$
the fluctuation of the integral does not exceed $\{\phi(\beta_n) - \phi(\alpha_n)\}\,W(\alpha_n, \beta_n)$, or is less than $\{\phi(b) - \phi(a)\}\,W(\alpha_n, \beta_n)$. Thus $W'(\alpha_n, \beta_n)$ is less than a fixed multiple of $W(\alpha_n, \beta_n)$. If then $\sum_{n=1}^{\infty} W(\alpha_n, \beta_n)$ is convergent, so also is
$$\sum_{n=1}^{\infty} W'(\alpha_n, \beta_n).$$

It now follows that a point of the perfect set P, at which $\sum_{n=1}^{\infty} W(\alpha_n, \beta_n)$ is convergent, is also a point of convergence of $\sum_{n=1}^{\infty} W'(\alpha_n, \beta_n)$. For the result obtained above may be applied to the part of P contained in a neighbourhood of the point, so chosen that, in it, $\Sigma W(\alpha_n, \beta_n)$ is convergent.

It has thus been shown that the set of points of P which are not points of convergence of $\Sigma W'(\alpha_n, \beta_n)$ is non-dense in P, since the corresponding condition for $\Sigma W(\alpha_n, \beta_n)$ is satisfied. Therefore the condition III, of § 464, holds for the function $f(x)\,\phi(x)$.

For the complete establishment of the theorem, there only remains to be proved that.

If $f(x)$ be summable in a perfect set P, contained in an interval (α, β), then

$$\Big[F(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F(x)\,d\phi(x)$$
$$= \sum_{n=1}^{\infty}\left\{\Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)\right\} + \int_{(P)} f(x)\,\phi(x)\,dx;$$

the summation being taken for all the intervals (α_n, β_n) contiguous to P, it being assumed that the sum is absolutely convergent.

This theorem can then be applied to the perfect set contained in any of the intervals contiguous to any one of the perfect sets $P_1, P_2, \ldots, P_\beta, \ldots$, employed in § 465, the conditions of the theorem being in each case satisfied.

Let $f(x)$ be expressed by the sum of the two functions $f_1(x)$, $f_2(x)$; where $f_1(x) = f(x)$, in P, and $f_2(x) = f(x)$ in all interior points of the intervals (α_n, β_n). Let $F_1(x) = \int_\alpha^x f_1(x)\,dx$, and $F_2(x) = \int_\alpha^x f_2(x)\,dx$; thus $F(x) = F_1(x) + F_2(x)$.

The L-integral $\int_\alpha^\beta f_1(x)\,\phi(x)\,dx$ exists, being equal to $\int_{(P)} f_1(x)\,\phi(x)\,dx$; also $\int_\alpha^\beta f_2(x)\,\phi(x)\,dx$ exists as a D-integral, being equal to

$$\Sigma \int_{\alpha_n}^{\beta_n} f_2(x)\,\phi(x)\,dx.$$

In each interval (α_n, β_n), $F_1(x)$ is constant, thus

$$\Big[F_2(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F_2(x)\,d\phi(x) = \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x).$$

Now $$\Big[F(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F(x)\,d\phi(x)$$

$$= \left\{\Big[F_1(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F_1(x)\,d\phi(x)\right\} + \left\{\Big[F_2(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F_2(x)\,d\phi(x)\right\};$$

the expression in the first bracket on the right hand side is equal to

$$\int_\alpha^\beta f_1(x)\,\phi(x)\,dx, \text{ or to } \int_{(P)} f_1(x)\,\phi(x)\,dx.$$

In order to prove the theorem, it is then sufficient to shew that

$$\Big[F_2(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F_2(x)\,d\phi(x) = \sum_{n=1}^{\infty}\left\{\Big[F_2(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F_2(x)\,d\phi(x)\right\};$$

this is the special case of the original theorem which arises when $f(x) = 0$, over the set P.

Consider the first m of the intervals (α_n, β_n); and let

$$(\gamma_1, \delta_1), (\gamma_2, \delta_2), \dots (\gamma_{m+1}, \delta_{m+1}), \text{ where } \gamma_1 = \alpha, \delta_{m+1} = \beta,$$

be the intervals complementary to the intervals $(\alpha_1, \beta_1), \dots (\alpha_m, \beta_m)$.

We have

$$\Big[F_2(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F_2(x)\,d\phi(x) - \sum_{n=1}^{n=m}\left\{\Big[F_2(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F_2(x)\,d\phi(x)\right\}$$

$$= \sum_{r=1}^{r=m+1}\Big[F_2(x)\,\phi(x)\Big]_{\gamma_r}^{\delta_r} - \sum_{r=1}^{r=m+1}\int_{\gamma_r}^{\delta_r} F_2(x)\,d\phi(x);$$

and the expression on the right-hand side may be written as

$$\sum_{r=1}^{r=m+1}\phi(\delta_r)\{F_2(\delta_r) - F_2(\gamma_r)\} + \sum_{r=1}^{r=m+1} F_2(\gamma_r)\{\phi(\delta_r) - \phi(\gamma_r)\}$$

$$- \sum_{r=1}^{r=m+1}\int_{\gamma_r}^{\delta_r} F_2(x)\,d\phi(x),$$

which is equivalent to

$$\sum_{r=1}^{r=m+1}\phi(\delta_r)\{F_2(\delta_r) - F_2(\gamma_r)\} + \int_\alpha^\beta \psi_m(x)\,d\phi(x),$$

where $\psi_m(x)$ is defined to be $F_2(\gamma_r) - F_2(x)$, in the intervals (γ_r, δ_r), and to be zero in the intervals $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \dots (\alpha_m, \beta_m)$.

The first term is numerically less than $\bar{\phi}\sum_{r=1}^{r=m+1}|F_2(\delta_r) - F_2(\gamma_r)|$, where $\bar{\phi}$ is the upper boundary of $|\phi(x)|$ in (α, β); and

$$\sum_{r=1}^{r=m+1}|F_2(\delta_r) - F_2(\gamma_r)|$$

does not exceed the sum of the fluctuations of $F_2(x)$ in the intervals

$$(\alpha_{m+1}, \beta_{m+1}), (\alpha_{m+2}, \beta_{m+2}), \dots,$$

contained in the intervals (γ_r, δ_r). It follows, from the convergence of the series of fluctuations in the intervals contiguous to P, that the first term is arbitrarily small, if m be sufficiently great; thus the first term is $< \epsilon_m$, where $\lim_{m\sim\infty} \epsilon_m = 0$.

If x is not a point of P, we have $\psi_m(x) = 0$, for all sufficiently large values of m; and if x is a point of P, it is in one of the intervals (γ_r, δ_r), where γ_r converges to x, as m is indefinitely increased, and thus $\psi_m(x)$

converges to zero. Since $\psi_m(x)$ is bounded, for all values of m and x, we have, by a theorem established in § 445,

$$\lim_{m\sim\infty}\int_\alpha^\beta \psi_m(x)\,d\phi(x) = \int_\alpha^\beta [\lim_{m\sim\infty}\psi_m(x)]\,d\phi(x) = 0.$$

It has now been shown that the limit, as $m \sim \infty$, of

$$\Big[F_2(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F_2(x)\,d\phi(x) - \sum_{n=1}^{n=m}\left\{\Big[F_2(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F_2(x)\,d\phi(x)\right\},$$

is zero. The theorem has therefore been established; and it has been proved that

$$\int_a^b f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_a^b - \int_a^b F(x)\,d\phi(x),$$

which is an extension, to the D-integral, of the theorem of § 445, for the L-integral.

475. As in § 445, it may be shown that the second mean value theorem is applicable to the integral of $f(x)\,\phi(x)$, where $f(x)$ is integrable (D), and $\phi(x)$ is monotone and bounded.

For $\int_a^b F(x)\,d\phi(x) = F(\mu)\{\phi(b) - \phi(a)\}$, where μ is some number in the interval (a, b); therefore

$$\int_a^b f(x)\,\phi(x)\,dx = F(b)\,\phi(b) - F(\mu)\{\phi(b) - \phi(a)\},$$

if we assume that $F(a) = 0$. This is equivalent to

$$\int_a^b f(x)\,\phi(x)\,dx = \phi(a)\int_a^\mu f(x)\,dx + \phi(b)\int_\mu^b f(x)\,dx.$$

The more general form of the theorem, including Bonnet's form, may be deduced, as in § 422. For the application to the representation of a function that is integrable (D) by trigonometrical or other series, the result here obtained is of decided importance.

EXTENSIONS OF THE DEFINITION OF THE DENJOY INTEGRAL

476. A generalization of the definition, given in § 464, of the Denjoy integral has been made by Denjoy* himself, by W. H. Young†, and by Khintchine‡, independently of each other. Instead of the series

$$\sum_{n=1}^{\infty} W(\alpha_n, \beta_n),$$

in which the terms are the fluctuations of $V(\alpha_n, x)$ in the intervals (α_n, β_n), the series $\sum_{n=1}^{\infty} | V(\alpha_n, \beta_n) |$ is employed. Thus, for the definition (4), in § 464, the following is substituted.

---

* *Annales sc. de l'école normale* (3), vol. XXXIII (1916), p. 127, and (3), vol. XXXIV (1917), p. 181. These memoirs are a continuation of researches in the *Journal de Math.* (7), vol. I (1915), p. 105, and *Bull. de la soc. Math. de France*, vol. XLIII (1915), p. 161.

† *Proc. Lond. math. Soc.* (2), vol. XVI (1916), p. 175.

‡ *Comptes Rendus*, Paris, vol. CLXII (1916), p. 287.

(4)′. *Let P be a perfect set of points in an interval (α, β), contained in (a, b), and let it be assumed that $f(x)$ is summable over the set P, and suppose that $V(\alpha', \beta')$ has been defined for every interval (α', β'), of (α, β), which contains no points of P as interior points. Let (α_n, β_n), where $n = 1, 2, 3, \ldots$, denote the intervals of (α, β) that are contiguous to P, and let it be assumed that the series $\sum_{n=1}^{\infty} | V(\alpha_n, \beta_n) |$ is convergent. Then $V(\alpha, \beta)$ is defined by*

$$V(\alpha, \beta) = \sum_{n=1}^{\infty} V(\alpha_n, \beta_n) + \int_{(P)} f(x)\, dx.$$

The definitions (1), (2), and (3), of § 464, are unaltered.

Instead of the condition III, of § 464, the following is substituted:

III′. *For every perfect set P, if $V(\alpha_n, \beta_n)$ have been calculated for every interval (α_n, β_n) contiguous to P, the set of points of P which are not points of convergence of $\sum_{n=1}^{\infty} | V(\alpha_n, \beta_n) |$ is non-dense in P.*

As before, a point of convergence of $\sum_{n=1}^{\infty} | V(\alpha_n, \beta_n) |$ is one which has a neighbourhood such that the series of those terms for which (α_n, β_n) is interior to the neighbourhood is convergent. Thus the neighbourhood contains a part of P, such that the series is convergent, when taken for the interval contiguous to that part of P. The conditions I, II, of § 464, remain unaltered.

It will be observed that, if the series $\sum_{n=1}^{\infty} W(\alpha_n, \beta_n)$ is convergent, so also is the series $\sum_{n=1}^{\infty} | V(\alpha_n, \beta_n) |$; but that the converse does not hold. It thus appears that the D-integral is a special class of the functions $V(a, x)$ which satisfy the wider set of conditions obtained by substituting (4)′ and III′ for (4) and III respectively.

When the function $f(x)$ is such that all the definitions (1), (2), (3), (4)′ and I, II, III′ are realized, the actual construction of $V(a, b)$ is carried out precisely in the same manner as in the case of the D integral, no essential change being necessary in the proof of existence of the integral.

477. As has already been pointed out in § 466, the conditions (3) and II contain a postulation that $V(a, x)$ shall be a continuous function of x, and it was shewn that the D-integral constructed in § 464 actually satisfies the requirement of this postulation. It will however be shewn that when the new conditions (4)′ and III′ are substituted for (4) and III, the postulation of continuity of $V(a, x)$, or of $F(x)$, is not satisfied unless the function $f(x)$ satisfies a certain condition.

The set H denotes the aggregate of the points at which $f(x)$ is not summable, and of the points in the neighbourhood of which $\Sigma\,|V(\alpha_n, \beta_n)|$ is not convergent.

The function $V(a, x)$ is clearly continuous at any point which does not belong to H, since $V(a, x+h) - V(a, x)$ is an L-integral, for sufficiently small values of $|h|$. Next, let G be a perfect set, contained in an interval (α, β), and assume that $f(x)$ is summable over G, and that every point of G is a point of convergence of the series $\Sigma\,|V(\alpha_n, \beta_n)|$, when the summation is taken for those intervals contained in (α, β) which are contiguous to G.

We have

$$V(a, x+h) - V(a, x) = \left[\int_{(G)} f(x)\,dx + \Sigma \int_{a_n}^{b_n} f(x)\,dx\right]_x^{x+h},$$

where (a_n, b_n) is contiguous to G. As $h \sim 0$, we see that

$$\left[\int_{(G)} f(x)\,dx\right]_x^{x+h} \sim 0,$$

on account of the continuity of an L-integral. In case x and $x+h$ are both points of G, $\left[\Sigma \int_{a_n}^{b_n} f(x)\,dx\right]_x^{x+h}$ also converges to zero, as $h \sim 0$, in such a manner that $x+h$ is always a point of G. But if $x+h$ is not restricted to be a point of G, it is in the interior of some interval (a_m, b_m), contiguous to G, and the sum $\left[\Sigma \int_{a_n}^{b_n} f(x)\,dx\right]_x^{x+h}$ contains the term

$$\int_{a_m}^{x+h} f(x)\,dx,$$

if h be positive, or $\int_{x+h}^{b_m} f(x)\,dx$, in case h be negative. It may happen that, as h converges in any manner to zero, $\int_{a_m}^{x+h} f(x)\,dx$, or $\int_{x+h}^{b_m} f(x)\,dx$, does not converge to zero, and in that case $V(a, x)$ would not be continuous at the point x, of G. We have $\left|\int_{a_m}^{x+h} f(x)\,dx\right| \leqq W_m$, where W_m is the upper boundary of $\left|\int_{a'}^{b'} f(x)\,dx\right|$ for all intervals (a', b') contained in (a_m, b_m). Let it be supposed that, for an infinite sequence of values of m, W_m is greater than some positive number k. However small h may be, there will be one of these intervals (a_r, b_r), for which $W_r > k$, in the interval $(x, x+h)$. Hence, in (a_r, b_r), there will be two points $a_r' = x + h_1$, and $b_r' = x + h_2$, for which the terms considered differ by at least k. If we take a sequence of values of h, such that $x+h$ approaches x along the sequence $\{a_r'\}$, we obtain a limit which differs by at least k from the limit obtained when $x+h$ has a sequence of values $\{b_r'\}$. It thus appears that $V(a, x+h) - V(a, x)$ could not, for both sequences of values of h, have the limit zero. Hence, with the hypothesis made, $V(a, x)$ would be discontinuous at the point x. In order that $V(a, x)$ may be continuous, it is therefore necessary and sufficient that $\lim\limits_{m \sim \infty} W_m = 0$. This condition must be satisfied for all the perfect sets contained in all the intervals contiguous to each of the sets $P_1, P_2, \dots P_\gamma$, employed in the construction given in § 464

It thus appears that the postulation contained in (3) and II, which is equivalent to the postulation that the integral $V(a, x)$ constructed, as in § 464, but with (4)′ and III′ substituted for (4) and III, should be continuous, is satisfied only if the upper boundary of $V(\alpha_n, x)$, in the interval (α_n, β_n), converges to zero, as $n \sim \infty$, for each of the perfect sets contained in any interval contiguous to P_γ, where P_γ is any one of the perfect sets employed in the construction of $V(a, b)$.

478. An integral $F(x) \equiv \int_a^x f(x)\,dx$, constructed in accordance with the system of definitions and postulations (1), (2), (3), (4)′, and I, II, III′, which include the condition that $F(x)$ is a continuous function, will be termed a *DKY*-integral, or *Denjoy-Khintchine Young integral*. It is clear that a *D*-integral is also a *DKY*-integral, but that the converse is not true.

It is shewn, as in § 472, that the sum theorem holds* also for the *DKY*-integral:

If $f^{(1)}(x), f^{(2)}(x)$ *have DKY-integrals in* (a, b), *then* $f^{(1)}(x) + f^{(2)}(x)$ *has also a DKY-integral in the interval, and*

$$\int_a^b \{f^{(1)}(x) + f^{(2)}(x)\}\,dx = \int_a^b f^{(1)}(x)\,dx + \int_a^b f^{(2)}(x)\,dx$$

The proof given in § 466 is applicable to shew that the set of those points of any given perfect set P, each of which is either a point of non-summability with respect to P, of the function $f(x)$, that is integrable (*DKY*), or else a point of non-convergence of $\Sigma\,|\,F(\beta_n) - F(\alpha_n)\,|$ taken for the intervals (α_n, β_n), contiguous to P, is non-dense in P.

The theorem that the *D*-integral $F(x)$ of a function $f(x)$, that is integrable (*D*), has almost everywhere a differential coefficient, equal to $f(x)$, does not hold in general for the *DKY*-integral. The corresponding theorem for the latter integral has been formulated by Denjoy†, who has introduced for the purpose the conception of an approximate differential coefficient; thus:

A continuous function $F(x)$ *possesses, at a point* x_0, *an approximate differential coefficient, finite, and equal to* $F'(x_0)$, *if the set of points* x *for which*

$$\left|\frac{F(x) - F(x_0)}{x - x_0} - F'(x_0)\right| < \epsilon,$$

---

* This is contrary to a statement of W. H. Young, *loc. cit.* pp. 208 and 211.

† *Annales sc. de l'école normale* (3), vol. XXXIII (1916), p. 170.

has the metric density unity at the point x_0, whatever value the positive number ϵ may have.

The property of the indefinite DKY-integral is then formulated as follows*.

If $F(x)$ be the integral $\int_a^x f(x)\,dx$, of a function $f(x)$ which is integrable (DKY) in the interval (a, b), $F(x)$ has at almost every point of (a, b), an approximate differential coefficient, or else an ordinary differential coefficient, equal to $f(x)$.

The necessary and sufficient conditions have been formulated by Khintchine (*loc. cit*), that the indefinite DKY-integral $F(x)$ should have almost everywhere a differential coefficient, equal to $f(x)$.

THE YOUNG INTEGRAL

479. The most general definition of a non-absolutely convergent integral has been given† by W. H. Young. His definition is obtained by relaxing the conditions from which it is inferred that the DKY-integral is continuous. Accordingly, the Young integral, or Y-integral, is not necessarily a continuous function of its upper limit. When it is in fact continuous, it is then a DKY-integral. When a is a limit of points of H on its right, and β is a limit of points of H on its left, the definition (3), and the assumption II, of § 464, are no longer taken to hold for all sets of intervals (α_n, β_n) such that $\alpha_n \sim \alpha$, $\beta_n \sim \beta$, but only when α_n and β_n are restricted, for all values of n, to be points of H, the set of points of non-summability of the function $f(x)$.

Thus the postulation II is replaced by the condition

$$\int_a^\beta f(x)\,dx = \underset{\alpha_n \sim \alpha,\, \beta_n \sim \beta}{[\lim]} \int_{\alpha_n}^{\beta_n} f(x)\,dx,$$

where the square bracket is taken to indicate that the limit does not necessarily exist unless all the points α_n, β_n belong to the set H. When α is not a limiting point of a part of H on its right, the points α_n will be unrestricted, as in the definition of the DKY-integral, and a similar remark holds as regards β. When neither α nor β is such a limiting point, the bracket may be removed from the limit, as it is then unrestricted.

An effect of the introduction of this restriction, as regards the mode of approach of the end-points of intervals to their limits, is that

$$F(x) \equiv \int_a^x f(x)\,dx$$

is not necessarily continuous at a point x which belongs to the set H. However, at such a point, $F(x)$ always belongs to the aggregate of limits of $F(x+h)$, or of $F(x-h)$, as $h \sim 0$.

---

* *Annales sc. de l'école normale* (3), vol. XXXIV (1917), p. 184.

† *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p. 175.

In order that $F(x)$ may be continuous in the whole interval (a, b), it is clearly necessary that the restricted limits and the unrestricted limits should in every case give the same result. Thus the Y-integral, when it is continuous, is necessarily a DKY-integral, but of course not necessarily a D-integral.

480. It will be shewn that the function $F(x)$ constructed, as in the case of the DKY-integral, or the D-integral, with the continuity condition modified in the manner explained above, has the following property

In any closed interval (α, β), *contained in* (a, b), $F(x)$ *assumes all values between its upper and lower boundaries in the interval* (α, β)

Let (γ, δ) be an interval contained in (α, β), and such as to contain a sequence of intervals (γ_n, δ_n), for which $\gamma_n \sim \gamma$, and $\delta_n \sim \delta$, and such that in each of the closed intervals (γ_n, δ_n) the condition is satisfied that $F(x)$ takes any value between its upper and lower boundaries in that interval.

Also let it be assumed that $F(\gamma)$, $F(\delta)$ are respectively the limits of $F(\gamma_n)$, $F(\delta_n)$. Let k be any number between the upper and lower boundaries of $F(x)$ in the closed interval (γ, δ). If n be sufficiently great, k is between the upper and lower boundaries of $F(x)$ in the interval (γ_n, δ_n), and is therefore the value of $F(x)$ at some point in that interval, and therefore in (γ, δ). Thus the theorem holds for the interval (γ, δ).

Commencing with the fact that the theorem holds for any interval (γ, δ) in which $f(x)$ is integrable (L), and following out the various stages of construction of $F(x)$ for the interval (α, β), as in § 465, we see, by continually applying the method of passing to the limit, that the theorem can be established for the interval (α, β).

481. The following extension of the theorem of § 474 will now be established:

If $f(x)$ *have a* Y*-integral* $F(x)$, *in* (a, b), *which is bounded in the interval, and* $\phi(x)$ *be a bounded monotone function defined in the same interval, then* $f(x)\phi(x)$ *is also integrable* (Y) *in* (a, b), *and*

$$\int_a^b f(x)\phi(x)\,dx = \Big[F(x)\phi(x)\Big]_a^b - \int_a^b F(x)\,d\phi(x).$$

It will, in the first place, be shewn that, if $F(x)$ be bounded in (α, β), and such that $\int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)$ exists, in accordance with the definition of § 445, for each of the intervals (α_n, β_n), of a sequence of intervals contained in (α, β), such that $\alpha_n \sim \alpha$, $\beta_n \sim \beta$, then $\int_\alpha^\beta F(x)\,d\phi(x)$ also exists.

Let $\xi = \phi(x)$, and suppose $\chi(\xi)$ to be the function defined as in § 445, for the interval $(\phi(\alpha), \phi(\beta))$ of ξ. By hypothesis, $\chi(\xi)$ is summable in each interval $(\phi(\alpha_n), \phi(\beta_n))$, of ξ. Let $E(A, n)$ denote the set of points of this interval for which $\chi(\xi) > A$, where A is an arbitrarily chosen number, then the set $E(A, n)$ is measurable. It then follows that the set $E(A)$, of all points ξ, of the interval $(\phi(\alpha + 0), \phi(\beta - 0))$, at which $\chi(\xi) > A$, is measurable, since it is the outer limiting set of the sequence of sets $E(A, n)$, as n is increased indefinitely. It follows that $\chi(\xi)$ is measurable in the interval $(\phi(\alpha + 0), \phi(\beta - 0))$; and since $\chi(\xi)$ is bounded, it is summable in the interval. Moreover, by a known property of the L-integral, we have

$$\int_{\phi(\alpha+0)}^{\phi(\beta-0)} \chi(\xi)\,d\xi = \lim_{n\sim\infty} \int_{\phi(\alpha_n)}^{\phi(\beta_n)} \chi(\xi)\,d\xi.$$

Since $\chi(\xi)$ has the value $F(\alpha)$ in the interval $(\phi(\alpha), \phi(\alpha+0))$, and the value $F(\beta)$ in the interval $(\phi(\beta-0), \phi(\beta))$, we have

$$\int_{\phi(\alpha)}^{\phi(\beta)} \chi(\xi)\,d\xi = F(\alpha)\{\phi(\alpha+0) - \phi(\alpha)\} + F(\beta)\{\phi(\beta) - \phi(\beta-0)\} \\ + \lim_{n\sim\infty} \int_{\phi(\alpha_n)}^{\phi(\beta_n)} \chi(\xi)\,d\xi;$$

therefore $\int_\alpha^\beta F(x)\,d\phi(x)$ exists, and has the value

$$F(\alpha)\{\phi(\alpha+0) - \phi(\alpha)\} + F(\beta)\{\phi(\beta) - \phi(\beta-0)\} + \lim_{n\sim\infty} \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x).$$

Next, it will be shewn that, if the theorem

$$\int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)$$

holds for each of the intervals (α_n, β_n), it also holds for the interval (α, β), provided $F(\alpha_n) \sim F(\alpha)$, $F(\beta_n) \sim F(\beta)$.

We have

$$\lim_{n\sim\infty} \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} = \Big[F(x)\,\phi(x)\Big]_{\alpha}^{\beta} - F(\beta)\{\phi(\beta) - \phi(\beta-0)\} \\ - F(\alpha)\{\phi(\alpha+0) - \phi(\alpha)\};$$

it follows that

$$\Big[F(x)\,\phi(x)\Big]_{\alpha}^{\beta} - \int_\alpha^\beta F(x)\,d\phi(x) = \lim_{n\sim\infty} \left\{ \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x) \right\}.$$

Since $\int_\alpha^\beta f(x)\,\phi(x)\,dx$ is defined to be $\lim_{n\sim\infty} \int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx$, it being assumed that $f(x)\,\phi(x)$ is integrable (Y) in each of the intervals (α_n, β_n), it follows that, if $\int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx$ exists, and is equal to

$$\Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)$$

for every value of n, then $\int_\alpha^\beta f(x)\,\phi(x)\,dx$ exists, and is equal to

$$\Big[F(x)\,\phi(x)\Big]_{\alpha}^{\beta} - \int_\alpha^\beta F(x)\,d\phi(x).$$

It must now be shewn that the condition III′ is satisfied by the integrals of $f(x)\,\phi(x)$, when it is assumed to be satisfied for the integrals of $f(x)$.

If (α_n, β_n) be an interval, contiguous to the perfect set P, in which the integral $f(x)\,\phi(x)$ has already been constructed, we have

$$\int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)$$

$$= \{F(\beta_n) - F(\alpha_n)\}\,\phi(\beta_n) + F(\alpha_n)\,\{\phi(\beta_n) - \phi(\alpha_n)\} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x).$$

If $\bar{F}$ denote the upper boundary of $F(x)$ in all the intervals (α_n, β_n) contiguous to P, and $\bar{\phi}$ is the upper boundary of $\phi(x)$ in (a, b), we have

$$\left|\int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx\right| \leq \bar{\phi}\,|F(\beta_n) - F(\alpha_n)| + 2\bar{F}\,\{\phi(\beta_n) - \phi(\alpha_n)\},$$

it being assumed that $\phi(x)$ is non-decreasing.

It follows that, if $\sum\limits_{n=1}^{\infty} |F(\beta_n) - F(\alpha_n)|$ is convergent, so also is

$$\sum_{n=1}^{\infty} \left|\int_{\alpha_n}^{\beta_n} f(x)\,\phi(x)\,dx\right|.$$

It is thence seen that, if the set of points of a perfect set which are points of divergence of the sum of the absolute values of the integrals of $f(x)$ taken over the intervals contiguous to the perfect set, is non-dense in the perfect set, the same condition holds as regards the absolute values of the integrals of $f(x)\,\phi(x)$. For, if P be the part of the perfect set contained in any interval (α, β), such that the sum of the absolute values of the integrals of $f(x)$ over the intervals contained in (α, β), contiguous to P, is convergent, it has been shewn that the sum of the integrals of $f(x)\,\phi(x)$ is absolutely convergent. Thus the condition III′ is satisfied for the function $f(x)\,\phi(x)$.

Lastly, it must be shewn that, if $f(x)$ be summable in a perfect set P contained in an interval (α, β), then

$$\Big[F(x)\,\phi(x)\Big]_{\alpha}^{\beta} - \int_{\alpha}^{\beta} F(x)\,d\phi(x) = \sum_{n=1}^{\infty}\left\{\Big[F(x)\,\phi(x)\Big]_{\alpha_n}^{\beta_n} - \int_{\alpha_n}^{\beta_n} F(x)\,d\phi(x)\right\} + \int_{(P)} f(x)\,\phi(x)\,dx;$$

the summation being taken for all the intervals (α_n, β_n) contiguous to P, it being assumed that the sum on the right-hand side is absolutely convergent. The proof of the corresponding theorem for the D-integral, given in § 474, is applicable to the present case.

We are now in a position to carry out all the stages of the construction of the Y-integral of $f(x)\,\phi(x)$, corresponding to the various stages in the construction of the Y-integral of $f(x)$. It has been shewn that the construction in the former case can always be carried out when it can be carried out in the latter case, and that at each stage the relation

$$\int_\alpha^\beta f(x)\,\phi(x)\,dx = \Big[F(x)\,\phi(x)\Big]_\alpha^\beta - \int_\alpha^\beta F(x)\,d\phi(x)$$

holds good The theorem is therefore completely established.

482. It has been shown that $F(x)$ assumes, in any interval, all the values between its upper and lower boundaries in that interval. Thus if U, L be the upper and lower boundaries of $F(x)$ in (a, b), we have

$$L\{\phi(b) - \phi(a)\} \leq \int_a^b F(x)\,d\phi(x) \leq U\{\phi(b) - \phi(a)\};$$

it follows that

$$\int_a^b F(x)\,d\phi(x) = F(\bar{x})\{\phi(b) - \phi(a)\},$$

where $\bar{x}$ is some number such that $a \leq \bar{x} \leq b$.

It now follows that

$$\int_a^b f(x)\,\phi(x)\,dx = \phi(a)\int_a^{\bar{x}} f(x)\,dx + \phi(b)\int_{\bar{x}}^b f(x)\,dx.$$

which is the second mean value theorem for the function $f(x)$, which is integrable (Y). The more general form of the theorem may be deduced, as in § 422.

It has been assumed throughout that $F(x)$ is bounded in the interval (a, b), and it is therefore subject to this condition that the existence of

$$\int_a^b f(x)\,\phi(x)\,dx,$$

as given by the formula for integration by parts, has been established.

It may however happen that, in the final stage of the construction of the Y-integral of a function $f(x)$, infinite discontinuities of the function $F(x)$ appear. Thus there exist Y-integrals which possess points of infinite discontinuity, and are thus unbounded.

It can be seen that the theorem

$$\int_a^b \{f_1(x) + f_2(x)\}\,dx = \int_a^b f_1(x)\,dx + \int_a^b f_2(x)\,dx$$

does not necessarily hold for Y-integrals, unless both the integrals on the right-hand side are DKY-integrals. If, for example, the set H, for $f_1(x)$, consists of the sequence $c_1, c_2, \ldots c_n, \ldots$ converging to zero, and the corresponding set for $f_2(x)$ consists of a different sequence $c_1', c_2', \ldots c_n', \ldots$ also converging to zero, $\int_0^1 f_1(x)\,dx$ is defined as $\lim\limits_{n\sim\infty}\int_{c_n}^1 f_1(x)\,dx$, and $\int_0^1 f_2(x)\,dx$ is defined as $\lim\limits_{n\sim\infty}\int_{c_n'}^1 f_2(x)\,dx$. But $\int_0^1 \{f_1(x) + f_2(x)\}\,dx$ would be defined as the limit of $\int_c^1 \{f_1(x) + f_2(x)\}\,dx$, where c has values belonging to both the sequences $\{c_n\}$, $\{c_n'\}$. Since it is not necessarily the case that $\int_{c_n'}^1 f_1(x)\,dx$ converges to $\int_0^1 f_1(x)\,dx$, or that $\int_{c_n}^1 f_2(x)\,dx$ converges to $\int_0^1 f_2(x)\,dx$, it follows that the sum theorem does not necessarily hold for Y-integrals which do not belong to the class of DKY-integrals.

EXAMPLES

1*. Let $f(x) = 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}$, for $0 < x \leq a$, and $f(0) = 0$. We have

$$f(x) = \frac{d}{dx}\left(x^2 \sin \frac{1}{x^2}\right),$$

for $0 < x \leq a$. $\int_\epsilon^a |f(x)|\,dx$ increases indefinitely as $\epsilon \sim 0$, and therefore $f(x)$ is not summable in the interval $(0, a)$. But $\int_\epsilon^a f(x)\,dx = \left[x^2 \sin \frac{1}{x^2}\right]_\epsilon^a$, and thus $\int_0^a f(x)\,dx$ exists as a D-integral, which is also an HL-integral.

2†. Let $$\phi_{\alpha\beta}(x) = \frac{d}{dx}\left[\frac{(x-\alpha)^2(\beta - x)^2}{(\beta-\alpha)^2} \sin \frac{k}{(x-\alpha)^m(\beta-x)^m}\right],$$

in the interval (α, β), where $m \geq 2$. The function $\phi_{\alpha\beta}(x)$ is not summable in the neighbourhoods of the points α, β; but $\int_\alpha^\beta \phi_{\alpha\beta}(x)\,dx$ exists as a D-integral, which is also an HL-integral, and it has the value zero. Let H be any non-dense closed set of points in the interval (a, b), and let $f(x)$ have the value zero at each point of H, and in each interval (α, β) contiguous to H, let it have the value $\phi_{\alpha\beta}(x)$. The integral $\int_\alpha^\beta \phi(x)\,dx$ exists, and has the value zero, in each of the intervals (α, β); also the fluctuation of $\int_\alpha^x f(x)\,dx$ in (α, β) is $< \frac{1}{6}(\beta - \alpha)^2$. It follows that the sum of the fluctuations of $\int_\alpha^x f(x)\,dx$ in all the intervals is absolutely convergent. It is now easily seen that, if G be the nucleus of H, the integral $\int_{\alpha_n}^{\beta_n} f(x)\,dx$ exists, and has the value zero, in every interval (α_n, β_n) contiguous to G; and moreover the sum of the fluctuations of the integrals $\int_{\alpha_n}^x f(x)\,dx$, in the intervals (α_n, β_n), is absolutely convergent. Therefore $f(x)$ has a D-integral, which has the value zero in any interval of which the end-points belong to H.

3‡. Let G be a perfect set, of measure zero, in the interval (a, b). In any interval (α_n, β_n), contiguous to G, let $f(x) = 1/(\beta_n - \alpha_n)$, for $\alpha_n < x < \frac{1}{2}(\alpha_n + \beta_n)$, and $f(x) = -1/(\beta_n - \alpha_n)$, for $\frac{1}{2}(\alpha_n + \beta_n) \leq x < \beta_n$; and let $f(x)$ have any set of values over G.

We have $\int_{\alpha_n}^{\beta_n} f(x)\,dx = 0$, for all values of n; also $\int_{\alpha_n}^{\frac{1}{2}(\alpha_n+\beta_n)} f(x)\,dx = \frac{1}{2}$. It thus appears that $f(x)$ is not integrable (D), because every point of G is a point of non-convergence of the sum of the fluctuations of the integral in the contiguous intervals. At the middle point of any interval (α_n, β_n), we have $F(x) = \frac{1}{2}$; hence, at any point of G which is a limiting point on both sides, we have

---

* See Denjoy, *Annales sc. de l'école normale* (3), vol. XXXIV (1917), p. 181.

† *Ibid.* p. 208.

‡ W. H. Young, *Proc. Lond. Math. Soc.* (2), vol. XVI (1916), p. 199.

$$F(x) = 0, \quad \overline{F}(x+0) = \tfrac{1}{2}, \quad \underline{F}(x+0) = 0, \quad \overline{F}(x-0) = \tfrac{1}{2}, \quad \underline{F}(x-0) = 0.$$

At a point of G which is an end-point of a contiguous interval, $F(x)$ is discontinuous on the side away from the contiguous interval The integral $F(x)$ is accordingly a Y-integral, which is not a DKY-integral.

4*. Let C denote Cantor's perfect set, of measure zero, defined in the interval $(0, 1)$ (see § 83, Ex. 1). In an interval (α_n, β_n), contiguous to C, let $f(x) = \dfrac{1}{m(\beta_n - \alpha_n)}$, for $\alpha_n < x < \frac{1}{2}(\alpha_n + \beta_n)$, and $f(x) = -\dfrac{1}{m(\beta_n - \alpha_n)}$, for $\frac{1}{2}(\alpha_n + \beta_n) \leqq x < \beta_n$; where m is such that $(\frac{1}{3})^m = \beta_n - \alpha_n$. Over the set C, $f(x)$ may have arbitrarily assigned values.

Since $\int_{\alpha_n}^{\beta_n} f(x)\,dx = 0$, we have $F(x) = 0$, at all points of C; and in (α_n, β_n), the maximum value of $\int_{\alpha_n}^{x} f(x)\,dx$ is $\dfrac{1}{2m}$.

The values of $\overline{F}(x+0)$, $\overline{F}(x-0)$, at a point x, of C, are the limits of the numbers $\dfrac{1}{2m}$, for the contiguous intervals in the neighbourhoods on the right and left of the point; and these limits are both zero, since m increases indefinitely with n; therefore $F(x)$ is a continuous function The integral is not a D-integral, for, if x be any point of C, there are contained in any neighbourhood of x, contiguous intervals for which m has all values which exceed some fixed value; and thus ΣW_n diverges at every such point x To see this, we observe that an integer r can be so determined that the distance of x from the nearer end-point of a given neighbourhood d, of x, is $> 3^{-r}$, and thus d will contain two consecutive divisions, when the whole interval $(0, 1)$ is divided into 3^{r+1} equal parts. We may suppose the intervals (α_n, β_n) to be arranged in descending order of their lengths, and such that those which are of equal length are arranged in their order from left to right The sub-set of C in any of the 3^{r+1} equal parts is similar to C, so that the values of m, for those contiguous intervals that are interior to d, will contain all integers from and after some fixed one.

Every point of G is a point of convergence of the sum of the absolute values of the integrals of $f(x)$ taken over all the contiguous intervals (α_n, β_n), since all these integrals have the value zero. It follows that $F(x)$ is a DKY-integral.

5† Let $f(x)$ be the function defined in Ex 4, and, denoting $\frac{1}{2}(\alpha_n + \beta_n)$ by γ_n, let $\phi(x)$ be defined in each interval (γ_n, β_n) by the same rule as that by which $f(x)$ was defined in the interval $(0, 1)$; and let $\phi(x) = 0$, at all points not in an interval (γ_n, β_n). The points of non-summability of $\phi(x)$ form the set $\overline{C}$, which consists of C, together with a set similar to C in each of the intervals (γ_n, β_n). It is clear that the set C is non-dense in the set $\overline{C}$. The points of non-summability of the function $f(x) + \phi(x)$ are the same as those of $\phi(x)$; that is they form the set $\overline{C}$.

We have $\int_{\alpha_n}^{\gamma_n} \{f(x) + \phi(x)\}\,dx = \int_{\alpha_n}^{\gamma_n} f(x)\,dx = \dfrac{1}{2m}$; and if $(\alpha_{ns}, \beta_{ns})$ denote those intervals contiguous to $\overline{C}$ that are in (γ_n, β_n), we have

$$\int_{\alpha_{ns}}^{\beta_{ns}} \{f(x) + \phi(x)\}\,dx = \int_{\alpha_{ns}}^{\beta_{ns}} f(x)\,dx = -\frac{(\beta_{ns} - \alpha_{ns})}{m(\beta_n - \alpha_n)} = -\frac{1}{2m \cdot 3^r},$$

where r depends on s, as m does on n.

---

* See W. H. Young, *loc. cit.* p 199

† See W. H. Young, *loc. cit* p 209. The result there given is not in accordance with that in the text, it is there stated that the series of integrals diverges in the neighbourhood of every point of $\overline{C}$, and thus that $f(x) + \phi(x)$ has no DKY-integral.

Those points of $\overline{C}$ that belong to C are points of divergence of the sum of the absolute values of the integrals of $f(x) + \phi(x)$, taken over the contiguous intervals (α_n, γ_n), $(\alpha_{ns}, \beta_{ns})$, because, in any interval δ which encloses such a point, there is an indefinitely great number of the intervals (α_n, γ_n), and the sum of the series of terms $1/2m$, taken for these intervals, diverges. This is however not the case for those points of $\overline{C}$ which do not belong to C, because such a point is interior to an interval (γ_n, β_n), or is at the end γ_n of such an interval, and the series $\Sigma \left| \int_{\alpha_{ns}}^{\beta_{ns}} \{f(x) + \phi(x)\}\, dx \right|$ is equal to $\frac{1}{m(\beta_n - \alpha_n)} \Sigma (\beta_{ns} - \alpha_{ns})$, which does not exceed $\frac{1}{2m}$. The set C being non-dense in $\overline{C}$, it now follows in accordance with the theorem of § 478, that the integral $\int_0^1 \{f(x) + \phi(x)\}\, dx$ exists as a DKY integral, and is equal to $\int_0^1 f(x)\, dx + \int_0^1 \phi(x)\, dx$, which has the value zero.

6*. Let G be a perfect set of content zero, in the interval (a, b), and let

$$f(x) = n/(b_n - a_n), \text{ for } a_n < x < \tfrac{1}{2}(a_n + b_n),$$
$$f(x) = -n/(b_n - a_n), \text{ for } \tfrac{1}{2}(a_n + b_n) < x < b_n,$$

where the intervals (a_n, b_n) are all contiguous to G. The function $f(x)$ may be defined arbitrarily in those points of G which are limiting points on both sides. The function $F(x)$ is zero at all the points of G, and it is continuous in each of the closed intervals (a_n, b_n). The value of $F(x)$ at the point $\frac{1}{2}(a_n + b_n)$ is $\frac{1}{2}n$. Thus at each point of G, $F(x)$ has an infinite discontinuity; at a point x which is a limiting point of G on both sides, we have

$$\overline{F}(x + 0) = +\infty, \quad \underline{F}(x + 0) = 0, \quad \overline{F}(x - 0) = 0, \quad \underline{F}(x - 0) = -\infty.$$

Thus $F(x)$ is a Y-integral, with a set of points of infinite discontinuity.

---

* See W. H. Young, *loc. cit.* p. 205.

CORRECTIONS AND ADDITIONS TO VOLUME II (1926)

Page 8. Ex. (7). This example should be stated as follows: If $\{\beta_n\}$ denote a monotone sequence of decreasing numbers which converges to 0, and $\{\alpha_n\}$ be a sequence which converges to 0, prove, by a method similar to that employed in the proof of the general theorem, that, if $\lim\limits_{n\sim\infty} \dfrac{\alpha_n - \alpha_{n+1}}{\beta_n - \beta_{n+1}}$ exists, then $\lim\limits_{n\sim\infty} \dfrac{\alpha_n}{\beta_n}$ exists, and has the same value

Page 249. A proof of the following more general theorem has been given by the author (*Journal of Lond. Math. Soc.* vol I (1926), p. 211).

Let E be a measurable set of points (x), *in any number of dimensions, and of either finite or infinite measure, and let* $\{f_n(x)\}$ *be a sequence of functions defined in E, and such that, for every value of n,* $|f_n(x)|^k$ *is summable over E, where k is some number greater than zero. If the condition*

$$\lim_{n\sim\infty,\ n'\sim\infty} \int_{(E)} |f_n(x) - f_{n'}(x)|^k\, dx = 0$$

be satisfied, then there exists a function $f(x)$, *defined at almost all points of E, such that* $|f(x)|^k$ *is summable over E, and satisfies the conditions*

$$\lim_{n\sim\infty} \int_{(E)} |f(x) - f_n(x)|^k\, dx = 0,$$

$$\int_{(E)} |f(x)|^k\, dx = \lim_{n\sim\infty} \int_{(E)} |f_n(x)|^k\, dx.$$

Page 290. It has been tacitly assumed here that the function $s(x)$ and the sets of points e_n are measurable In order to justify these assumptions, we observe that, in accordance with I, § 400, $s(x)$ is a measurable function, and consequently (see I, § 384) the functions $s(x) - s_{n+m}(x)$ are also measurable, hence the set of points at which $|s(x) - s_{n+m}(x)| \leq \epsilon$ is measurable, for each value of $n+m$. The set e_n is the set of points common to all the sets for which

$$|s(x) - s_n(x)| \leq \epsilon,\quad |s(x) - s_{n+1}(x)| \leq \epsilon,\ \ldots\ |s(x) - s_{n+m}(x)| \leq \epsilon,\ \ldots$$

are measurable. In accordance with the second theorem in I, § 131, the set e_n is measurable

Page 323. It is here assumed that the sets e_h, e_{-1} are measurable It has however been pointed out to the author by R L. Jeffery, of Acadia University, Nova Scotia, that $f(x, y)$ can be so defined that this is not the case For example, let E denote the set of all points of the interval $(0, 1)$, and let L denote a non-measurable linear set of points, to which the point 0 may be taken to belong, which is non-measurable in each interval $(0, h)$ Such a set has been constructed by Van Vleck (*Trans. Amer. Math. Soc.*, vol. IX (1908), p. 237). Let $f(x, y)$ be defined in the rectangle $(0, 0, 1, 1)$ to have the value 0 at every point in the rectangle, except those points on the diagonal (x, x) which have as their projections on the side $(0, 1)$ points of $C(L)$, at which $f(x, x) = 1$. For all points x on $(0, 1)$, we have $\lim\limits_{y\sim 0} f(x, y) = 0$; moreover, $f(x, y)$ is measurable for each value of y. The set e_h consists of the interval $(h+0, 1)$ together with the points of L in the interval $(0, h)$; thus e_h is not measurable in the interval $(0, 1)$.

Whenever this possibility is realized it is necessary in § 225 to restrict h to have the values in an enumerable sequence $\{h_n\}$ which converges to zero, the values of y being also restricted to have the enumerable set of values of $y_0 + h_n$.

The sets $e_{h_1}, e_{h_2}, \ldots$ are then measurable, as has been shewn in the last note. Since then the theorem holds for all such sequences $\{h_n\}$, it holds in accordance with Heine's definition of the limit, and consequently in accordance with the definition of Cauchy, provided the equivalence of the two definitions be assumed (see I, § 211); this assumption requires the employment of the multiplicative axiom. The case of the sets e_1 can be treated similarly. The argument of the preceding note is not applicable to shew that the set of points common to an unenumerable family of measurable sets is measurable.

Line 1. *For* $y+h$ *read* y_0+h.

Line 7. *For* $\lim\limits_{h\sim 0} m(E-e_h)$ *read* $\overline{\lim\limits_{h\sim 0}}\, m(E-e_h)$.

Line 11. *For* $|f(\xi, y_0+h_n) - f(x, y_0+0)|$ *read* $|f(\xi, y_0+h_n) - f(\xi, y_0+0)|$.

Line 12. *For* $f(x, y_0+0)$ *read* $f(\xi, y_0+0)$.

Page 444. Insert, after line 4, the following.

Further, $\int_a^{x-\mu} F(x', x, n)\,dx'$ *converges to zero, as* $n\sim\infty$, *uniformly for all those values of* x *(in* G*) that are* $\geq \alpha+\mu$, *and* $\int_{x+\mu}^{\beta} F(x', x, n)\,dx'$ *converges to zero, uniformly for all values of* x *(in* G*)* $\leq \beta-\mu$.

The following theorem, which is an immediate corollary of Theorem I, is usually sufficient for application.

If $\Phi(x', x, n)$ *has the value* $F(x', x, n)$ *for all values of* x' *such that* $|x'-x| \leq \mu$, *and is zero whenever* $|x'-x| > \mu$, *and if* $\Phi(x', x, n)$ *so defined, satisfies the conditions* (1) *and* (2), *of Theorem I, then*

$$\int_a^{x-\mu} F(x', x, n)\,dx' + \int_{x+\mu}^{b} F(x', x, n)\,dx'$$

converges to zero, uniformly for all values of x *(in* G*).*

Page 453. Line 2. *For* "the integral is" *read* "the integral differs from

$$\frac{\chi(\alpha_n)}{\alpha_n}\int_{\alpha_n}^{\mu} t^2 F_1(t, n)\,dt \text{ by.''}$$

Line 6. *For* $N(\alpha_n)\, V_{\alpha_n}^{\mu}\left\{\frac{\chi(t)}{t}\right\}$ *read* $N(\alpha_n)\left[V_{\alpha_n}^{\mu}\left\{\frac{\chi(t)}{t}\right\} + \frac{\chi_1(\alpha_n)}{\alpha_n}\right]$.

Line 7 from the foot. *For* $\alpha_n V_{\alpha_n}^{\mu}\left\{\frac{\chi(t)}{t}\right\}$ *read* $\alpha_n V_{\alpha_n}^{\mu}\left\{\frac{\chi(t)}{t}\right\} + \chi_1(\alpha_n)$.

Page 558. Line 13 from the foot. *For*

$\lim\limits_{n\sim\infty}\int_0^{\frac{1}{2}\pi} 2\left[\tfrac{1}{2} + \cos 2t + \cos 4t + \ldots + \cos 2nt\right]dt$ *read* $\lim\limits_{n\sim\infty}\int_0^{\frac{1}{2}\pi} 2\left(\frac{1}{2} + \sum_{r=1}^{n-1}\frac{n-r}{n}\cos 2rt\right)dt$.

Page 626. The theorem in § 409 has been extended by Kolmogoroff and Seliverstoff to the case $\epsilon = 0$ (*Atti dei Lincei, Rendiconti* (6), vol. II (1926), p. 307). Thus, *if* $\sum\limits_{n=2}(a_n^2+b_n^2)\log n$ *is convergent, then* $\sum\limits_{n=1}(a_n \cos nx + b_n \sin nx)$ *converges almost everywhere*. An independent proof of this result has been given by Plessner (*Crelle's Journal*, vol. CLV (1926), p. 15).

Page 700. Delete lines 1–11. It is not correct that $\frac{1}{t^{(1)}t^{(2)}} - \frac{1}{\sin t^{(1)} \sin t^{(2)}}$ is summable over the cell $(0, 0; \frac{1}{2}\pi, \frac{1}{2}\pi)$. The two following corrections are requisite in consequence of this.

Page 705. § 464, line 5. *For* $\int F(t^{(1)}, t^{(2)}) \frac{\sin m^{(1)}t^{(1)}}{t^{(1)}} \frac{\sin m^{(2)}t^{(2)}}{t^{(2)}}\, d(t^{(1)}, t^{(2)})$ *read*

$$\int\left[\frac{t^{(1)}t^{(2)}}{\sin t^{(1)} \sin t^{(2)}} F(t^{(1)}, t^{(2)})\right]\frac{\sin m^{(1)}t^{(1)}}{t^{(1)}} \frac{\sin m^{(2)}t^{(2)}}{t^{(2)}}\, d(t^{(1)}, t^{(2)}).$$

Page 709. Delete the last two lines, and substitute the following:

The functions F_1, F_2 can be so chosen as to be not only non-increasing, but also non-negative. The function $\frac{t^{(1)}t^{(2)}}{\sin t^{(1)}\sin t^{(2)}}$ can, by proper choice of A, be expressed as the difference of A and $A-\frac{t^{(1)}t^{(2)}}{\sin t^{(1)}\sin t^{(2)}}$, both of which are non-negative and non-increasing. The function $\psi(t^{(1)}, t^{(2)})$ can be made identical successively with each of the four functions

$$AF_1(t^{(1)}, t^{(2)}), \quad \left(A-\frac{t^{(1)}t^{(2)}}{\sin t^{(1)}\sin t^{(2)}}\right)F_1(t^{(1)}, t^{(2)}),$$
$$AF_2(t^{(1)}, t^{(2)}), \quad \left(A-\frac{t^{(1)}t^{(2)}}{\sin t^{(1)}\sin t^{(2)}}\right)F_2(t^{(1)}, t^{(2)}),$$

all of which are monotone non-increasing.

The limits of the four corresponding integrals, as $m^{(1)}\sim\infty$, $m^{(2)}\sim\infty$, are $\frac14\pi^2AF_1(+0,+0)$, $\frac14\pi^2(A-1)F_1(+0,-0)$, $\frac14\pi^2AF_2(-0,-0)$, $\frac14\pi^2(A-1)F_2(-0,+0)$ respectively.

Thus the integral

$$\frac{1}{\pi^2}\int_{(0,0)}^{(\frac12\pi,\frac12\pi)}F(t^{(1)}, t^{(2)})\frac{\sin m^{(1)}t^{(1)}}{\sin t^{(1)}}\frac{\sin m^{(2)}t^{(2)}}{\sin t^{(2)}}\,d(t^{(1)}, t^{(2)})$$

converges to $\frac14F(+0,+0)$, as $m^{(1)}\sim\infty$, $m^{(2)}\sim\infty$.

Page 768. The first theorem in § 495 holds good, as asserted by Menchoff, when $2>\lambda>0$. It is correct that the condition $2>\lambda>\frac12$ is necessary in order that the first theorem in § 495 may be deducible from the second. But, for the second theorem, the following more general one may be stated, it is proved on pp. 778, 779. From this more general theorem the more general form of the first can be deduced:

If $\omega(n)$ is a positive increasing function, such that the series $\sum_{p=1}^{\infty}\frac{1}{\omega(p)}$ is convergent, and if $a_n=o(1)$ the convergence of the series

$$\sum_{n=1}^{\infty}\left[\omega\left(\log\log\frac{1}{|a_n|}\right)\right]^2\left(\log\frac{1}{|a_n|}\right)^2a_n^2 \quad \ldots\ldots\ldots \quad \text{A}$$

entails the convergence, almost everywhere in the interval (0, 1), of the series

$$\sum_{n=1}^{\infty}a_n\phi_n(x).$$

For, let $\omega(n)=n^2$, then $\sum_{p=1}^{\infty}\frac{1}{\omega(p)}$ is convergent, also from the convergence of $|a_n|^{2-\lambda}$, we have $a_n=o(1)$.

We have also

$$\lim_{u\sim\infty}\frac{[\omega(\log\log u)]^2(\log u)^2}{u^\lambda}=\lim_{u\sim\infty}\frac{(\log\log u)^4(\log u)^2}{u^\lambda}=0,$$

for $\lambda>0$. Hence, for all sufficiently large values of n,

$$\left[\omega\left(\log\log\frac{1}{|a_n|}\right)\right]^2\left(\log\frac{1}{|a_n|}\right)^2<C\frac{1}{|a_n|^\lambda},$$

where C is a positive constant, thus, when $\lambda>0$

$$\left[\omega\left(\log\log\frac{1}{|a_n|}\right)\right]^2\left(\log\frac{1}{|a_n|}\right)^2a_n^2<C|a_n|^{2-\lambda}.$$

Therefore the convergence of the series $\Sigma|a_n|^{2-\lambda}$ entails the convergence of the series (A), and consequently the convergence almost everywhere, of the series $\Sigma a_n\phi_n(x)$.

LIST OF AUTHORS QUOTED IN VOLUME I

[*The numbers refer to pages*]

Ampère, 354
Archimedes, 57
Arzelà, 185, 343, 514
Ascoli, 55, 74, 462

Baire, 77, 120, 134, 158, 263, 307, 311, 312, 449, 450
Baker, H. F., 111
Banach, 400
Beman, 10
Bendixson, 117, 127
Bernstein, 209, 213, 224, 240, 244, 246, 247, 252, 260, 267
Bettazzi, 53, 319
Biermann, 22
Bliss, 664
Bolza, 157, 436
Bolzano, 62
Bonnet, 618
Borel, 76, 106, 110, 111, 141, 161, 178, 180, 185, 209, 212, 256, 263, 289, 573
Brodén, 124, 138, 139, 275, 277, 388, 389, 402
Bromwich, 353
Brouwer, 144, 158
Burali-Forti, 259
Burkill, 401, 704

Cantor, G., 10, 22, 31, 32, 49, 50, 52, 74, 75, 76, 77, 81, 83, 84, 85, 87, 91, 97, 98, 101, 121, 123, 128, 161, 163, 201, 202, 211, 237, 238, 240, 242, 251, 255, 259
Cantor, M., 20, 37, 368
Carathéodory, 161, 186
Cathcart, 421
Cauchy, 459, 461, 493
Chwistek, 252
Couturat, 10

Dantscher, 445
Darboux, 462
Dedekind, 10, 21, 22, 56
Denjoy, 136, 138, 156, 196, 312, 400, 401, 692, 699, 715, 718, 724
Dini, 302, 317, 356, 366, 386, 390, 426, 432, 443, 468, 469
Dirichlet, 274, 297, 320, 459
Dixon, A. C., 639
Du Bois-Reymond, 38, 39, 55, 74, 75, 77, 117, 355, 382, 391, 473, 492, 496, 510, 518, 618, 619

Ettlinger, 517
Euclid, 55
Euler, 49

Faber, 83
Fichtenholz, 293, 653
Fourier, 273
Fraenkel, 252
Frege, 1, 11
Fubini, 426, 609, 631

Genocchi, 405
Gillespie, D. C., 517, 551
Goursat, 111
Gravé, 368
Gross, 106

Hadamard, 263
Hahn, 324, 366, 485, 573, 594, 670, 672
Hamilton, W. R., 11
Hankel, 22, 161, 317, 320, 619
Hardy, G. H., 242, 247, 345, 491, 530, 546, 559
Harnack, 117, 161, 163, 170, 317, 321, 421, 459, 514, 657
Hartog, 269
Hausdorff, 158, 247, 252
Hedrick, 443
Heine, 10, 22, 31, 111, 282, 290
Hellinger, 669
Helmholtz, 2, 10, 11
Hermite, 84
Hessenberg, 252
Hilbert, 40, 251, 455, 458
Hildebrandt, 573, 663, 664, 674
Hobson, 90, 144, 157, 409, 424, 426, 436, 443, 582, 619, 631, 632, 644, 672, 674, 688
Holder, 53, 55, 619
Huntington, 224, 251
Husserl, 11

Jackson, Dunham, 293
Jacob, S. M., 21
Jordan, 76, 146, 147, 197, 325, 338, 443, 462, 476, 514, 519, 527, 657

Jourdain, 213, 226, 238, 251, 252, 259, 267

Khintchine, 715, 719
Konig, J., 90, 242, 252, 377
Kopcke, 390
Korselt, 209
Kowalewski, 619
Krause, 345
Kronecker, 22, 619
Küstermann, 346

Lambert, 49
Lebesgue, 161, 166, 191, 195, 263, 287, 336, 337, 338, 386, 400, 455, 459, 465, 508, 572, 573, 581, 588, 596, 606, 623, 624, 636, 637, 641, 643, 649, 664, 683
Legendre, 49
Leibnitz, 1
Levi, B., 267, 582, 596
Lichtenstein, 517
Lindelof, 110, 129, 130
Lindemann, 84
Liouville, 84, 124
Locke, 1
Looman, 312, 692
Loria, 451
Luroth, 356, 386, 469
Lusin, 112, 196

Mahlo, 136
Martinotti, 409
Méray, 22
Meyer, 619
Mill, J. S., 1
Mirimanoff, 260, 262
Mittag-Leffler, 75, 99, 100
Mollerup, 251
Montel, 400
Moore, E. H., 457, 458, 674, 676, 680

Nalli, 692
Netto, 451
Neumann, C., 351, 619
Neville, 80
Newton, 20

Osgood, 443

Padoa, 8
Pal, 114
Pasch, 302
Peacock, 21
Peano, 10, 24, 197, 263, 321, 338, 405, 426, 429, 445, 452, 458
Pereno, 390
Pierpont, 514, 533, 573
Pincherle, 22
Poincaré, 8, 263, 269
Pollard, 166, 546, 548, 557
Prasad, G., 485
Pringsheim, 10, 20, 49, 50, 200, 356, 360, 498, 504, 514, 517, 518, 618, 619, 648

Rademacher, 432
Radon, 674
Rajchman, 400
Richard, 90
Rider, 409
Riemann, 320, 459, 461, 464, 504, 648
Riesz, F., 216, 595, 643, 645, 674
Rudio, 49
Russell, B. A. W., 10, 11, 24, 31, 99, 224, 251, 254, 261
Ruziewicz, 366

Saks, 409
Scheeffer, 338, 375, 366, 368, 382, 445
Schepp, 58
Schoenflies, 117, 118, 138, 158, 213, 251, 252, 260, 321, 324, 340, 377, 396, 457, 458, 476, 485, 496, 510, 516
Schroder, 209
Schubert, 11
Schwarz, 370, 426
Serret, 514
Sierpiński, 98, 130, 196, 246, 266, 282, 632
Singh, A. N., 401
Smith, H. J. S., 75, 123, 476
Steinitz, 390
Stéphanos, 50
Stieltjes, 538
Stolz, 55, 57, 161, 360, 424, 426, 428, 446, 476, 518, 519
Study, 331, 338

Tannery, 32
Thomae, 404, 420, 462, 476, 494, 517
Tonelli, 609

Vallée Poussin, de la, 70, 78, 79, 108, 155, 168, 173, 177, 178, 191, 193, 196, 293, 459, 466, 568, 586, 590, 603, 616, 653
Van Vleck, 166, 664
Veblen, 224
Veltmann, 200
Vergerio, 345
Veronese, 53, 58
Vitali, 186, 582, 614

Volterra, 278, 320, 490

Wallis, J., 37
Watson, G. N., 158
Weierstrass, 22, 62, 354, 618
Westfall, 443
Whitehead, 11, 90, 206, 252
Whittaker, J. M., 555

Young, G. C., 114, 116, 158, 383, 384, 391, 393. 400, 401, 450
Young, R. C., 415
Young, W. H., 104, 105, 106, 111. 114. 116, 117, 128, 130, 139, 158, 174, 183. 185, 241, 304, 314, 316, 327, 345, 347, 360, 383, 384, 400. 402, 405, 416, 426, 427. 432. 434. 443, 450, 473, 516, 542, 546, 560, 572, 594. 600, 609. 625. 650, 666, 668, 691, 715, 718. 719, 724, 725, 726

Zermelo, 209, 251, 262, 263, 268, 269

GENERAL INDEX TO VOLUME I

[*The numbers refer to pages*]

Adherence, 131
Aggregate(s); General notion of, 1, Elements of, 1; Finite, 3, Similar, 4, 217, 255, Enumerable, 81, Equivalent, 9, 254. Rank of elements in, 6; Ordinal number of, 4, 230, Perfect, 52, 220; Straight line as, 59, Power of, 86, Definition of, 202, 250, Comparable and Incomparable, 205, 268; Closed, 220, Dense in itself, 220; Everywhere dense, 220, General theory of, 250, Normally ordered, 225, Parts of, 204; Limiting elements of, 219; Principal element of, 219, of order-types, η, θ, π, 221, Normally ordered, 225, Segment of normally ordered, 226; Simply ordered, 216, 253, Structure of simply ordered, 219
Aleph numbers, 207, General theory of, 236, 258
Archimedean system, 42
Archimedes; Axiom of, 57
Arithmetization, 14; Kronecker's scheme of, 22

Cardinal number(s); Definition of, 9, 203, 251, Addition and Multiplication of, 205, as exponents, 206, Division of, 213, exceeded by other cardinal numbers, 211, of continuum, 242, 246, of second class of ordinals, 235, Relative order of, 204; Smallest transfinite, 207; of aggregate of continuous functions, 289, of aggregate of all functions, 289
Categories, First and Second, 131
Coherence, 131
Congruency, Axiom of, 55
Connexity of aggregate of real numbers, 51
Content, of sets of points, 161, 169
Continuity, of functions, 281; Uniform, 290, of unbounded functions, 293; Absolute, 291; Approximate, 312
Continuous functions, 283; defined for a continuous interval, 286; two definitions not equivalent, 287; defined at points of a set, 287; in an open interval, 305, Oscillating, 374, 386; Construction of, 387
Continuum, of real numbers, 51; given by intuition, 54; Straight line as, 55; Arithmetic, 52, 88; Cardinal number of, 86, 242, 246; of p-dimensions, 87; Order-type of, 223
Convergence, General principle of, 38
Correspondence, General notion of, 2
Curves filling space, 155

Derivatives, of sets of points, 74, 81, Transfinite, 99, of functions, 352, Upper and lower, 354, Progressive and Regressive, 354; Bounded, 375, Properties of, 365, 380, 385; General properties of, 391
Differential Coefficients, 352, 356, Successive, 368, Partial, 417, Higher partial, 422
Discontinuity, Infinite, 280, of functions, 300; Ordinary, 301, Classification, 301, of first and second kinds, 301
Discontinuous functions, Classification of, 313, Point-wise, 314, 316, Definition of, by extension, 322

Enumerable aggregate, Definition, 81, of rational numbers, 83, of algebraical numbers, 83, isolated sets, 102
Equivalence Theorem, 209

Fluctuation, of functions, 280, Inner, 280
Fractional numbers, 14
Frontier of set of points, 78, 116
Functional image, 273
Functional limits, Aggregate of, 298; Symmetry of, 304
Functional relation, 272, 277
Functions, Dirichlet's definition, 271; Homonomically and heteronomically defined, 275; of a variable aggregate, 277, Upper and lower boundaries and limits of, 278; Bounded, 279; Unbounded, 279; Upper boundary of, in an interval, 279, Continuity of, 281, Continuity of unbounded, 293; Limits of, at a point, 295; Maximum and Minimum of, at a point, 297, 300; Semi continuous, 307; Approximately continuous, 312; Classification of discontinuous, 313; Point-wise discontinuous, 316; of bounded variation, 325, 343; Monotone, 318, 343; Quasi-monotone, 347; of bounded total fluctuation, 331; Most nearly continuous, 324; Maxima and Minima and lines of invariability of, 348; Derivative of, 352; Differential coefficients of, 356, with lines of invariability, 367; Successive differential coefficients of, 368; Oscillating continuous, 374; with bounded derivatives, 375;

Properties of derivatives of continuous, 380, with one derivative assigned, 386, Ordinary, 391, Construction of continuous, 387, Measurable, 395, 562; of two variables, 404; Double and repeated limits of, 405, Existence and equality of repeated limits of, 408 Limits of monotone, of two variables, 414; defined implicitly, 432, Maxima and Minima of, of two variables, 444, linear in each interval of a set, 491

Generation, Cantor's principle of, 93
Graphs, Functions defined by, 391

Heine-Borel theorem, 106

Incrementary ratios, 377, Boundaries of, 382
Indeterminate forms, Evaluation of, 359
Inequalities involving integrals, 642
Infinitesimals, Non-existence of, 42, Veronese's theory of, 58
Integers, Sets of sequences of, 158
Integrable functions, Particular cases of, 468, null-functions, 480
Integral Calculus, Fundamental theorem of the, 159, 482, 596, 699
Integral(s), Riemann, 460, Upper and lower Riemann, 462, Conditions of existence of, 465, Properties of definite Riemann, 472; Riemann of functions of two or more variables, 476, Cauchy's definition of improper, 493; Riemann, over unbounded interval, 498, Change of variable in a single Riemann, 506, Repeated Riemann, 509; Improper double, 518, Double, over an infinite domain, 527, Transformation of double, 531, Riemann-Stieltjes, 538, 666, Upper and lower Generalized Riemann-Stieltjes, 546, 547; Riemann-Stieltjes, for functions of two variables, 559; Lebesgue, of a measurable function, 565, Other definitions of an, 572; Lebesgue, as measure of a set of points, 573, the Riemann, as a Lebesgue integral, 575; Lebesgue, as function of a set of points, 576, Equivalent Lebesgue, 578, Properties of Lebesgue, 579, Indefinite, 587, Upper and lower semi-, 594, Total variation of an indefinite, 595; Generalized indefinite, 607; Indefinite, of a function of two variables, 608; Repeated Lebesgue, 626; Approximate representation of, as a Riemann sum, 640; Lebesgue, over an infinite field, 646; Change of independent variable in a Lebesgue, 649, Harnack's definition of an, 657, Lebesgue-Stieltjes, 662; Hellinger's, 669; Non-absolutely convergent, 675; Harnack-Lebesgue, 676; Denjoy, 692; Denjoy-Young-Khintchine, 715, Young, 719
Integration; Geometrical interpretation of Riemann, 470, by parts, 491, 616, 691, 711
Intervals, Open and Closed, 52, Sets of, 100; Lebesgue chain of, 114
Irrational numbers, 20; Dedekind's theory of, 23, Cantor's theory of, 28

Limits; Arithmetical theory of, 37; Method of, 38
Lines of invariability, 348

Magnitude, Theory of, 20, 55
Maxima and Minima, 300, 348; of a function of two variables, 444
Mean value theorem, of Differential Calculus, 357, of Integral Calculus, 617, 618, 624, 689, 715, 723
Measure; Problem of, 165; of open and closed sets, 167, Exterior and Interior, 172
Measurable sets of points, 174; Related to a system of sets, 176, 182, (B), 179, (J), 197
Measurement, 53
Metric density, 190
Metrically dense, 191
Metrical properties, 196

Negative numbers, 18
Nets, Systems of, 68
Null-function, Point-wise discontinuous, 323

Order, Notion of, 2
Order-functions, 240
Order-type(s), 216, Addition and Multiplication of, 217; of continuum, 222; η, θ, π, 221
Ordinal numbers; Definition of, 4, 230; Transfinite, 91; of second class, 94, 232; Limiting, 232; Arithmetic of, 238; Sum and product of, 230
Oscillation, of function at a point, 301; of function in an interval, 350

Rational numbers; Aggregate of, 19, Section of aggregate of, 23
Real numbers; Dedekind's definition, 24; Cantor's definition, 30; Operations on, 32; Convergent sequences of, 35; Equivalence of definitions of, 40; Representation of, 47; everywhere dense, 51
Rectifiable curves, 338

Repeated limits; of functions, 405; Existence and equality of, 408; of integrals, 509, 626

Saltus, 300, -function, 308
Schwarz's theorem, 370, on partial differential coefficients, 426
Semi-continuous functions, 307
Sequences, Simple, 6; Convergent, 28; Ascending and Descending, 219; Double, 571, Monotone double, 582
Sets of intervals or cells; Limiting point of, 65, Properties of, 100; External, internal, and semi-external point of, 104
Sets of points, 46, Analysis of, 128, Closed, 76, 97, 116, 125, 146, Connex, 51, 148, Degrees of points in, 128; dense in themselves, 76; Derivatives of, 74, 81, Detached, 147, Everywhere dense, 77, Frontier of, 78, 146, Improper limiting point of, 72, Inner and outer limiting, 133, Interior points of, 77; Irreducible, 125; Isolated, 76; Limiting points of, 70; Measurable, 174; Non-dense, 77, Non-dense perfect, 117, of first and second species, 74; of first and second categories, 134, Open, 78, 149; Perfect, 76, Reducible, 125, Residual, 136; Sequences of closed, 97, Sequences of, 183, Transfinite derivatives of, 98, Unbounded, 62 Measure of unbounded, 184, Upper and lower boundaries of, 61, of cardinal number of the continuum, 246
Square, Representation of, in linear interval, 151
Systems of Nets, 68

Total fluctuation, 331

Unenumerable, Proof that continuum is, 84
Unity; Notion of, 1

Variable, Real, 271
Variation, 325, 341
Vitali's theorem, 186

CAMBRIDGE: PRINTED BY WALTER LEWIS, M.A., AT THE UNIVERSITY PRESS